METHODS IN PRODUCTIVITY AND EFFICIENCY ANALYSIS WITH APPLICATIONS TO WAREHOUSING

A Thesis
Presented to
The Academic Faculty

by

Andrew Johnson

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May, 2006

METHODS IN PRODUCTIVITY AND EFFICIENCY ANALYSIS

WITH APPLICATIONS TO WAREHOUSING

Approved by:

Dr. Leon F. McGinnis, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Steven T. Hackman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Paul M. Griffin
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Leonard Parsons
College of Management
Georgia Institute of Technology

Dr. Gunter P. Sharp
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: March 31, 2006

To Chiaki

ACKNOWLEDGEMENTS

I would like to express my sincerest thanks and gratitude to my advisor, Dr. Leon McGinnis. It was his initial encouragement that began my doctorial research and it has been his patience and support that has helped to continue my research efforts. I would also like to thank Dr. Paul Griffin for his assistance and frequent conversations. It is his open door which allows so many students to connect with this department and feel a genuine concern and interest of the faculty. I would further like to thank Dr. Guntar Sharp, Dr. Steve Hackman, and Dr. Len Parsons for having served on my committee and for their valuable advice. I also owe Dr. Knox Lovell many thanks for his time and lessons on production theory; it is with his insight and discussions my understanding of these topics has become much richer. This dissertation would not have been possible without all the assistance I have received.

Special thanks to Dr. Wen-chih Chen for his assistance and advice in helping me develop tools involved in this research. I would also like to acknowledge the lab mates that have worked side-by-side with me, including but not limited to Yen-Tai Wan, Jinxiang Gu, Dima Nazzal, Maga Khachatryan, and Uanny Brens-Gracia, who have each made working in the lab much more interesting. In particular I would like to thank Gonzalo Cordova who has challenged me both mentally with our discussion about industrial engineering and physically in the student recreational center.

I thank my parents for their unconditional love and support on all the decisions I have made. Without them none of this would have been possible. And most importantly I

would like to express my sincere gratitude to my wife, Chiaki; it is with her by my side that life is more enjoyable.

Funding for the research has come from the National Science Foundation, W.M. Keck Foundation grant to the Virtual Factory Lab, from The Logistics Institute through the membership of the Progress Group, and from Georgia Tech through the Gwaltney Chair for Manufacturing Systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation addresses a set of technical issues related to benchmarking best practice behavior in warehouses. In order to identify best practice, first performance needs to be measured. A variety of tools are available to measure productivity and efficiency. One of the most common tools is data envelopment analysis (DEA). Given a population of systems that consumes inputs to generate outputs, production theory can be used to develop basic postulates about the production possibility space and to construct an efficient frontier which is used to quantify efficiency for individual systems. Beyond inputs and outputs, warehouses typically have practices (techniques used in the warehouse) or attributes (characteristics of the environment of the warehouse including demand characteristics) which also influence efficiency. Previously in the literature, a two-stage method has been developed to investigate the impact of practices and attributes on efficiency. When applying this method, two issues arose: how to measure efficiency in small samples and how to identify outliers. The small sample efficiency measurement method developed in this thesis is called multi-input / multi-output quantile based approach (MQBA) and uses deleted residuals to estimate efficiency. The outlier detection method introduces the inefficient frontier. Both overly efficient and overly inefficient outliers can be identified by constructing an efficient and an inefficient frontier. The outlier detection method incorporates an iterative procedure previously described, but has not implemented in the literature. Further, this thesis also discusses issues related to selecting an orientation in super efficiency models. Super efficiency models are used in outlier detection, but are also commonly used in measuring technical

progress via the Malmquist index. These issues are addressed using two data sets recently collected in the warehousing industry. The first data set consists of 390 observations of various types of warehouses. The other data set has 25 observations from a specific industry. For both data sets, it is shown that significantly different results are realized if the methods suggested in this document are adopted.

CHAPTER 1

INTRODUCTION

As *industrial engineers* we are "concerned with the design, improvement, and installation of integrated systems of people, material, information, equipment, and energy. We draw upon specialized knowledge and skills in the mathematical, physical, and social sciences together with the principles and methods of engineering analysis and design to specify, predict, and evaluate the results to be obtained from such systems" Institute of Industrial Engineers [2005]. The industrial engineer needs to have knowledge of the domain to predict and evaluate the results of various operational and design decisions.

When systems are designed, and operating methods and parameters are specified, the industrial engineer often would like to quantify the effects of various choices in terms of changes in the production possibilities and the system's efficiency. The *production function* gives a mathematical representation of the trade-off between resource inputs and production outputs. If an industrial engineer had the production function, the cost for every design, the costs of inputs, and price of outputs, it would be a straightforward mathematical exercise to find the profit maximizing design. Hence, the ability to develop the production function or an approximation to it would have a major impact on design decision making.

An alternative to the production function is the production possibility set. This set consists of pairs of input and output vectors, such that the input vector can produce the output vector. Non-parametric approaches make assumptions about ways the production

possibility set can be constructed and from this production possibility set the same information as contained in the production function can be inferred.

This research has developed from the desire to use non-parametric estimates of the production function to evaluate warehouses. In the process of applying these techniques, it has become apparent that the methodology for computing non-parametric estimates of the production function has not been completely developed. There are significant issues yet to be resolved. This research attempts to resolve some of these issues and uses the new methods developed to evaluate distribution centers.

1.1 Measuring productivity in a warehouse

A *warehouse or distribution center* is a location, typically occupying a large building and using specialized equipment to store goods temporarily, in the process of supplying those goods to customers or other members of a supply chain. To measure productivity within a distribution center, first the activities of a distribution center need to be defined. A particularly hard-to-handle activity within some distribution centers is value-added processing. When these activities are considered, the line between warehousing and manufacturing becomes blurred. Hence for the research presented here, the boundary of the distribution center system will be specified such that value-added activities are outside the scope of a distribution center.

The activities that define a distribution center are unloading, inspecting, putting away, storage, picking, grouping, packing and loading. The inputs consumed and the outputs generated from these activities need to be identified and quantified in order to measure the productivity of the distribution center. Activities that may be co-located with the

distribution center, but do not fall within the activities stated will be excluded from the analysis. Examples include inside sales and housekeeping.

There are many factors not under the control of the distribution center that affect efficiency. Some examples could be weather, market conditions, and other companies' competitive behavior. These variables cannot be controlled by the warehouse management for either short-term or long-term decision making. There also are decisions that can be made in long-term planning, which cannot be changed when making short term decisions. Some examples are facility location, role within the supply-chain, and what technologies to adopt. Figure 1.1 below lists attributes, inputs and outputs for both the short-run and long-run problems.

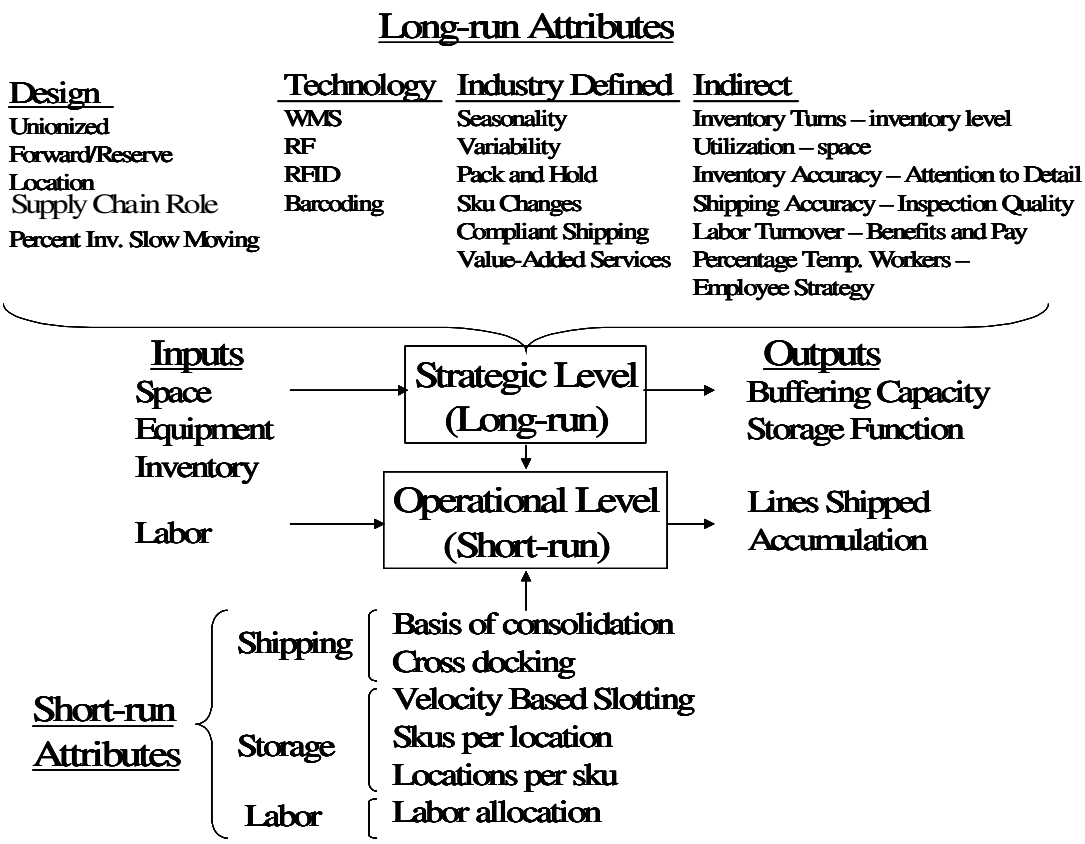

Figure 1.1: Identification of inputs, outputs and attributes for long-run and short-run attributes problems

A challenge in measuring warehouse efficiency is to identify the inputs, outputs and attributes which most influence the warehouse's efficiency. However, warehouses are not all alike, and various sub-groups can be identified. For example warehouses can be grouped by industry; e.g. automotive, electronics, pharmaceuticals, etc. The most appropriate inputs, outputs and attributes may be different for each industry. It is important to pick a production function model that is appropriate for the group being studied. While information can be used about the industry or group to identify a model, it may also be possible to identify observations from the subgroup which are not consistent with the model and the group of warehouses identified for analysis. This process of

identifying observations that are inconsistent with the rest of the data is referred to as outlier detection. From the results of an outlier detection method, an analyst can identify and correct data entry errors or observations that do not belong to the group. This allows a warehouse to be compared to a group of warehouses to which it is truly similar and makes benchmarking results and best practice results more accurate.

Figure 1.1. also identifies a variety of attributes which affect efficiency, but are neither inputs or outputs. The two-stage method was developed to investigate how these attributes affect efficiency. Efficiency estimates are calculated in the first stage and the estimates are regressed against the attributes in the second stage. From the second stage results it can be determined if the attribute has a significant correlation with efficiency.

The traditional non-parametric models used to measure efficiency typically make an assumption either that all inputs levels can be adjusted (this is called an input orientation) or assume all outputs levels can be adjusted (this is called an output orientation). However, when the purpose of the model is to quantify the distance from an observation to the rest of the data set, as in outlier detection, it is possible that using either the input or the output orientation may not be appropriate. Thus an alternative approach to these traditional orientations is desirable.

When the number of warehouses being analyzed is large, there is reason to believe the sample is representative of the population, and that estimates of the production function are reasonably accurate. However in smaller data sets it is more difficult to identify the production function for the entire group which is largely unobserved. Thus in small data sets, non-parametric methods may not work well. Because our goal is to understand the effects of operational and design decisions, it will suffice to estimate

relative efficiency when the data set is small. From a set of relative efficiency estimates, the best performance among the observed warehouses still can be identified.

1.2 Problem statement

The goal of this research is to develop new methods that support the estimating of productivity and using it for quantitative benchmarking. To this end, observations which are dissimilar to the group under evaluation should be identified to determine if they truly are members of the group to be evaluated. This problem has been described for the situation when only efficiency estimates are being calculated; however, the outlier problem has not been addressed for the two-stage method, which will be further described in chapter 2.

The orientation decisions for super efficiency models are an important aspect of using them in practice. The super efficiency model measures the efficiency of an observation relative to a production possibility set constructed from a set of observations which does not include the observation under analysis. Super-efficiency models are used in outlier detection but also are important in the calculation of Malmquist productivity indices. It has been shown previously in the academic literature that there are cases for both the input and output orientations in which the linear programs for computing a component of the Malmquist index have no feasible solution. In the context of outlier detection, this implies cases in which the decision to flag an observation would be made with little or no information. As a remedy to this problem the hyperbolic oriented super efficiency model is suggested and investigated in Chapter 5.

The methods described previously to quantify efficiency depend on identifying the efficient frontier. This can be difficult when not enough data is available to characterize

the sample distribution or identify the efficient frontier. Thus the quantile-based method for efficiency quantification is presented and developed. This method quantifies the relative efficiency of observations without identifying the efficient frontier and presents itself as an alternative in small data sets.

The issues described in this subsection were identified in the course of attempting to apply nonparametric methods to the analysis of warehouse performance data collected over a period of approximately five years. Two data sets were developed – one large data set representing a cross section of all warehouses, and a smaller, industry-specific dataset. The large data set contains almost 400 records, while the small data set contains only 25 records. These two data sets are used as case studies to demonstrate the methods developed in this thesis.

1.3 Organization of the thesis

Chapter 2 will review the basic mathematics of production theory. Chapter 3 will review the literature pertinent to the problems being addressed. Chapter 4 will develop methods for identifying outlier data using a "super efficiency" approach while Chapter 5 will establish the benefits of using a non-traditional efficiency measure in these super efficiency models. In Chapter 6 a quantile-based method for efficiency evaluation in small samples will be introduced and discussed, and in Chapter 7 the methods presented in previous chapters will be demonstrated in a case study examining distribution centers. Finally, Chapter 8 will summarize conclusions and describe possible extensions.

CHAPTER 2

METHODS TO BE USED

2.1 Defining terminology

The words decision making unit (DMU), observation and firm will be used interchangeably to mean a member of the group for which efficiency is being quantified. *Input* of a DMU is human, financial, or physical resources put into a system in order to achieve a result. *Output* of a DMU is any result, product, or service that a system produces. A production set is a collection T of pairs of vectors (x, y), where $x = (x_1, ..., x_i, ...x_P)$ is a vector of quantities of P inputs and $y = (y_1, ..., y_i, ...y_Q)$ a vector of quantities of Q outputs, which has the property of being *feasible*. By 'feasible' is meant that the output bundle y physically can be produced by making use of the input bundle x. In formal terms,

$$T = \{(x, y) \,|\, x \in R_+^P, y \in R_+^Q; (\text{x,y}) \text{ is feasible}\} \tag{2.1}$$

Given a set of N DMUs, the $N \times P$ input matrix, X, and the $N \times Q$ output matrix, Y, represent the input and the output data for all N firms.

Productivity expresses a ratio of outputs produced to inputs consumed. In the multi-input or multi-output case, productivity can be expressed either as a set of ratios or as a single ratio, in which case a method for aggregating inputs and outputs must be specified. *Efficiency* is a measure of performance relative to some reference value. The difference between productivity and efficiency is that productivity for an individual DMU can be

calculated without reference to any other DMU, whereas efficiency is a ratio of productivities with reference to an efficient frontier.

An *efficient frontier* is a description of the correspondence between input and output bundles when a DMU is operating at the "best case" productivity level.

Two types of efficient frontiers exist: the unobservable frontier based on the technology available, called the *feasible efficient frontier*, and the frontier constructed from observable instances of production, called the *observable efficient frontier*. When an industry is not building new facilities or adapting new technologies as they are developed, these two frontiers can differ significantly. Ideally, when measuring efficiency, the feasible efficient frontier should be used; typically the observable efficient frontier is used due to lack of information about the feasible efficient frontier. Observable efficient frontiers will be the focus of this work; thus, any reference to the efficient frontier will mean the observable efficient frontier, unless otherwise stated.

An important property of production functions is the concept of returns to scale. There are three types of returns to scale: constant, increasing and decreasing. To define returns to scale, assume we have a production possibility set T. Now suppose a DMU is using some vector of inputs x and some vector of outputs y. Consider scaling all inputs up or down by some amount $t \geq 0$. A technology exhibits constant returns to scale if any of the following are satisfied:

(1) y in T implies ty is in T, for all $t \geq 0$;

(2) x in $V(y)$ implies tx is in $V(ty)$ for all $t \geq 0$ (where $V(\ )$ is an inverse production function);

(3) $f(ty) = t\,f(y)$ for all $t \geq 0$; i.e., the production function $f(x)$ is homogeneous of degree 1.

A technology exhibits increasing returns to scale if $f(tx) > t\,f(x)$ for all $t > 1$.

A technology exhibits decreasing returns to scale if $f(tx) < t\,f(x)$ for all $t > 1$.

Since the seminal paper of Farrell [1957] a variety of models have emerged to measure what Farrell identified as technical inefficiency. Farrell defined technical efficiency based on a two input, one output model. Technical inefficiency was defined as a ratio of the distance from the origin (or reference point) to the efficient frontier over the distance from the origin (or reference point) to the point being measured (both measured along the same ray). This defines how the measure is constructed, but to make technical inefficiency a usable concept it needs to be related to the characteristics of the DMU under comparison.

While some people assume technical efficiency and managerial efficiency can be used interchangeable, Farrell gives the following description of technical efficiency,

> "Technical efficiency, then, is defined in relation to a given set of firms, in respect of a given set of factors measured in a specific way, and any change in these specifications will affect the measure. This is inevitable in any such measure. But with these qualifications it functions in a natural and satisfactory way as a measure of efficiency…

This statement implies that for technical inefficiency, as calculated by Farrell's method, to be related to actual unobservable value of efficiency for a DMU under evaluation, it is necessary to have a complete set of peer DMUs to evaluate against, the factors selected in the model as inputs and outputs must be correctly identified, and there should be no error in the measurement of the inputs and outputs. If any of these requirements are not met, the estimate of technical efficiency will be biased. Farrell goes on to say,

> … a firm's technical efficiency will reflect the quality of its inputs as well as the efficiency of its management. If these differences in quality are physically measurable, it may be possible to reduce this effect by defining a larger number of relatively homogeneous factors of production, but in practice it is never likely to be possible completely to eliminate it…
> …it is impossible to measure the efficiency of its management entirely separately from this factor (quality of inputs). This is, indeed, as it should be, for it is never possible to decide precisely how far the fertility of a particular farmer's land is due to nature and how far to good husbandry, how far the laziness and intractability of a particular firm's labour force is ingrained and how far the product of bad management."

While it is clear many variables (such as input quality) can be argued to be under managerial control, in a system there clearly are many variables beyond the manager's control. Some examples of these are seasonality of demand, location, and required response time.

Closely related to Farrell's working is the work of Shephard. A valuable concept in measuring efficiency is the distance function developed by Shephard [1970]. Here, if given the representation of a production technology as

$$L(y) = \{x : (y, x) \text{ is feasible}\} \tag{2.2}$$

which for every $y \in R_+^Q$ has isoquant

$$IsoqL(y) = \{x : x \in L(y),\ \lambda x \notin L(y), \lambda \in [0,1]\} \tag{2.3}$$

and efficient subset

$$EffL(y) = \{x : x \in L(y),\ x' \notin L(y), x' \leq x\} \tag{2.4}$$

then Shephard's input distance function provides a functional representation of a multiple input / multiple output technology and can be stated as

$$D_I(y, x) = \max\{\delta : (x / \delta) \in L(y)\} \tag{2.5}$$

Combining these results we can see for an inefficient DMU, $D_I(y,x) > 1$. Now if we represent a Farrell input oriented measure of technical efficiency as

$$F_I(y,x) = \min\{\theta : \theta x \in L(y)\} \tag{2.6}$$

it follows from duality

$$F_I(y,x) = \frac{1}{D_I(y,x)} \tag{2.7}$$

This presentation has taken an input orientation, but similar arguments hold and similar measures can be defined for an output orientation.

2.1.1 Typical production function

A typical production function relates the inputs consumed to the output generated with an allowance for error in measurement or random behavior. This can be written as

$$y = f(x,\beta) + \varepsilon \tag{2.8}$$

where y is a single output, x is vector of inputs, β are parameters estimated based on an initial data set, and ε are the residual error terms. The function $f(\ )$ can take on a variety of forms. The three most common are:

1) Cobb-Douglas, which has the form $y = x_1^{\alpha} x_2^{1-\alpha}$ where α is related to the rate of substitution and $0 \le \alpha \le 1$,

2) constant elasticity of substitution (CES) which has the form $y = \left[a_1 x_1^{\rho} + a_2 x_2^{\rho} \right]^{\frac{1}{\rho}}$ where a_1 and a_2 are weights defining the relative importance of the inputs and ρ is the elasticity of substitution $(0 \le \rho \le 1)$, and

3) translog cost (c) function which has the form

$\log c(w, y) = a_0 + \sum_{i=1}^{k} a_i \log w_i + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} b_{ij} \log w_i \log w_j + \log y$ where w_j is the price of input j, $\sum_{i=1}^{k} a_i = 1$, $\sum_{j=1}^{k} b_{ij} = 0$,and $b_{ij} = b_{ji}$.

The translog is perhaps the most popular because it allows a great deal of flexibility in the relationship between inputs and outputs. Often second order terms for the individual inputs are included and interaction terms between two inputs are included. While this increases flexibility there are still certain shapes that are not achievable with the relationships that are enforced through the selection of the production functions. This motivates the implicit production function used in data envelopment analysis and other linear programming based efficiency measurement methods to be described later in this chapter. While linear programming is generally considered to allow a more flexible representation of the production function, this is not always true; and there are efficient frontier shapes associated with translog production functions for instance that can not be obtained by linear programming, Lovell [1993].

2.1.2 Defining methods

Many methods have built upon the previous definitions, including stochastic production frontier approach (SFA), data envelopment analysis (DEA), deterministic frontier approach (DFA), individual specific effects (ISE) and free disposal hull (FDH), Bauer et al. [1998]. DEA is a nonparametric approach to measuring efficiency which uses linear programming techniques. One linear program needs to be generated and solved to calculate the efficiency score of each DMU. DMU s are identified for which no other DMU or linear combination of firms can produce as much or more of every output

(given an input level for all inputs) or use as little or less of every input (given an output level for all outputs).

The DEA efficient frontier is composed of these undominated DMUs and the piecewise linear segments which connect the set of input/output combinations of these DMUs, yielding a convex production possibilities set. The efficient frontier is defined by certain convex combinations of these undominated DMUs; since these composite DMU do not have an observable instance, they create composite DMUs with composite levels of input and output. These composite DMUs are called *virtual producers*.

The linear program decides the weighting of the efficient DMUs to construct a virtual DMU for the purposes of determining the efficiency of the DMU under evaluation. If the virtual DMU is better than the DMU being evaluated by either making more output with the same or less input or making the same output with less input, then the evaluated DMU is inefficient. Take for example the case where a virtual producer can make the same output with less input than DMU A. It is then said a proportional contraction of all resources, also called an equaproportional contraction, can occur. The size of this contraction (call this b) relative to the distance function measured to the point representing DMU A (call this a), can be used to calculate the efficiency of unit A by the equation $1-{}^{b}\!/_{a}$.

A fundamental assumption behind DEA and the use of virtual producers is a composite producer can be constructed by operating parts of a new producer unit in the manner of observed producers. If this is not true, then the virtual DMUs do not correspond to DMUs that could exist. Also a necessary assumption is that, if a given DMU, is capable of producing output level Y with input level X, then other producers

in the data set should also be able to do the same if they were to operate efficiently. If this assumption does not hold, then the set of producers under evaluation may not truly be peers.

There are two common variations of DEA used, constant returns to scale (CRS) and variable returns to scale (VRS). The difference in the linear programming formulation is a single constraint, often referred to as the convexity constraint. This constraint forces the weights assigned to construct the virtual unit to sum to one. This precludes a very small DMU from being scaled up several times with a weight greater than one, and forces the virtual DMU to be composed of at least one DMU producing more output than the DMU under evaluation. A desirable attribute of DEA is that it does not require the explicit specification of a functional form and so imposes very little structure on the shape of the efficient frontier.

Free disposal hull (FDH), like DEA is a nonparametric method for evaluating efficiency of individual producers based on a comparison set of data. For the same set of data, DEA's constructed production possibility set contains FDH's production possibility set. Linear programming techniques can be use to solve for efficiency estimates using a FDH model. However, a min/max formulation of FDH solves much faster than the linear programming counterpart. By definition FDH is the smallest free disposal set containing all observations in a sample of producers. Free disposal implies if a DMU producing a particular level of output y with a particular level of input x, was given more of any input the producer could freely give away or destroy the extra input and still produce the same level of output. DEA also makes this assumption, however, the distinction between DEA and FDH is that DEA also assumes convex combinations of any observed

production possibilities can be achieved whereas FDH does not. Therefore, FDH production possibility set is typically non-convex.

Parametric methods such as stochastic frontier approach (SFA) and individual specific effects (ISE) impose more structure on the shape of the frontier by specifying a functional form for the production function. As noted in Farrell [1957] the production function is never known in practice, thus the function can be estimated from sample data using either a nonparametric piece-wise-linear technology or a parametric function, such as the Cobb-Douglas form.

2.2 Data Envelopment Analysis

DEA was first introduced in ratio form, Charnes, Cooper and Rhodes [1978], which lends itself to an easy understanding, so we will follow that method here. For each DMU a measure of the ratio of weighted outputs over weighted inputs is calculated, e.g. $u'y_i \big/ v'x_i$, where u is a $Q\times 1$ vector of output weights and v is a $P\times 1$ vector of input weights and u' is the transpose of u. Consider the situation in which various DMUs place different levels of importance on particular inputs and outputs. Then the weights u and v would become specific to each DMU. DEA selects, for each DMU, the set of weights that maximizes the estimate of that DMU's efficiency by solving

$$\begin{aligned} &\max_{u,v}(u'y_i / v'x_i), \\ &s.t.\ \ u'y_j / v'x_j \le 1,\ j=1,2,...,N, \\ &u,v \ge 0 \end{aligned} \tag{2.9}$$

Then $u'y_i \big/ v'x_i$ is an efficiency measure for the ith DMU. The mathematical program maximizes efficiency of the ith DMU subject to the constraint all efficiency measures

must be less than or equal to one. This model is then solved once for each DMU. One problem with this ratio formulation is it has an infinite number of solutions. For example if (u,v) is a feasible solution then $(\alpha u, \alpha v)$ is also a solution. One way to avoid this is to impose an arbitrary value on the sum of one of the vectors or the sum of the product of the vector and another vector. Traditionally, $v'x_i = 1$ is chosen which allows the model to be rewritten in the linear form:

$$\begin{aligned} &\max_{u,v}(u'y_i), \\ &s.t.\ \ v'x_i = 1, \\ &\qquad u'y_j \le v'x_j,\ j = 1,2,...,N, \\ &\qquad u,v \ge 0 \end{aligned} \tag{2.10}$$

The resulting linear program is often referred to as the multiplier form of DEA.

The linear programming dual of (2.10) is called the envelopment problem:

$$\begin{aligned} &\min_{\theta_I,\lambda}(\theta_I), \\ &s.t.\ \ -y_i + Y\lambda \ge 0, \\ &\qquad \theta_I x_i - X\lambda \ge 0, \\ &\qquad \lambda \ge 0 \end{aligned} \tag{2.11}$$

given a set of N DMUs, where θ is a scalar and λ is a $N \times 1$ vector of variables. The $N \times P$ input matrix, X, and the $N \times Q$ output matrix, Y, represent the input and the output data for all N firms. This formulation finds a weighting of the N DMUs used in the analysis for which the composite DMU produces at least as much output as the DMU under evaluation without using more inputs than the minimal equaproportional reduction of the evaluated DMU's inputs. Because θ_I is associated with the input constraints, this model is often referred to as the input oriented model. Similarly, there is an output-oriented model

$$\begin{aligned} &\max_{\theta_O,\lambda}(\theta_O), \\ &s.t.\ \ -\theta_O y_i + Y\lambda \geq 0, \\ &\qquad x_i - X\lambda \geq 0, \\ &\qquad \lambda \geq 0 \end{aligned} \tag{2.12}$$

The output oriented formulation finds a weighting of the N DMUs used in the analysis for which the composite DMU uses no more of any input then the DMU under evaluation while producing more outputs than the maximal equaproportional increase of the evaluated DMU's outputs. Under a constant returns to scale assumption $\theta_I = 1/\theta_O$.

The frontier constructed in DEA is based on the following postulates described in Banker, Charnes and Cooper [1984]:

Postulate 1 (Convexity). If $(x_1, y_1) \in T$ and $(x_2, y_2) \in T$, then for any scalar $\theta \in [0,1]$, $(\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta) y_2) \in T$.

Postulate 2 (Monotonicity). (a) If $(x, y) \in T$ and $x_1 \geq x$, then $(x_1, y) \in T$.
(b) If $(x, y) \in T$ and $y_1 \leq y$, then $(x, y_1) \in T$.

Postulate 3 (Inclusion). The observed $(x_j, y_j) \in T$ for all DMUs $j = 1,...,n$.

Postulate 4 (Minimum extrapolation). If a production possibility set T_1 satisfies Postulates 1,2, and 3 above, then $T_1 \subseteq T$.

To calculate a DEA efficiency estimate two distances are compared. Efficiency is calculated as the ratio of the distance to the frontier divided by the distance to the point whose efficiency is being measured. With efficiency measured in this way input oriented DEA always gives efficiency estimates less than or equal to 1, while output oriented DEA gives efficiency estimates greater than or equal to 1. By examining figure 2.1, the input efficiency estimate can be described as holding the output level constant and

comparing $D_f^{input} / D_a^{input}$. Similarly, the output efficiency estimate can be seen to be the ratio of $D_f^{output} / D_a^{output}$. When the model becomes larger than two inputs and a single output (or any combination of inputs and outputs summing to three) these distances can no longer be visualized, but they can be calculated.

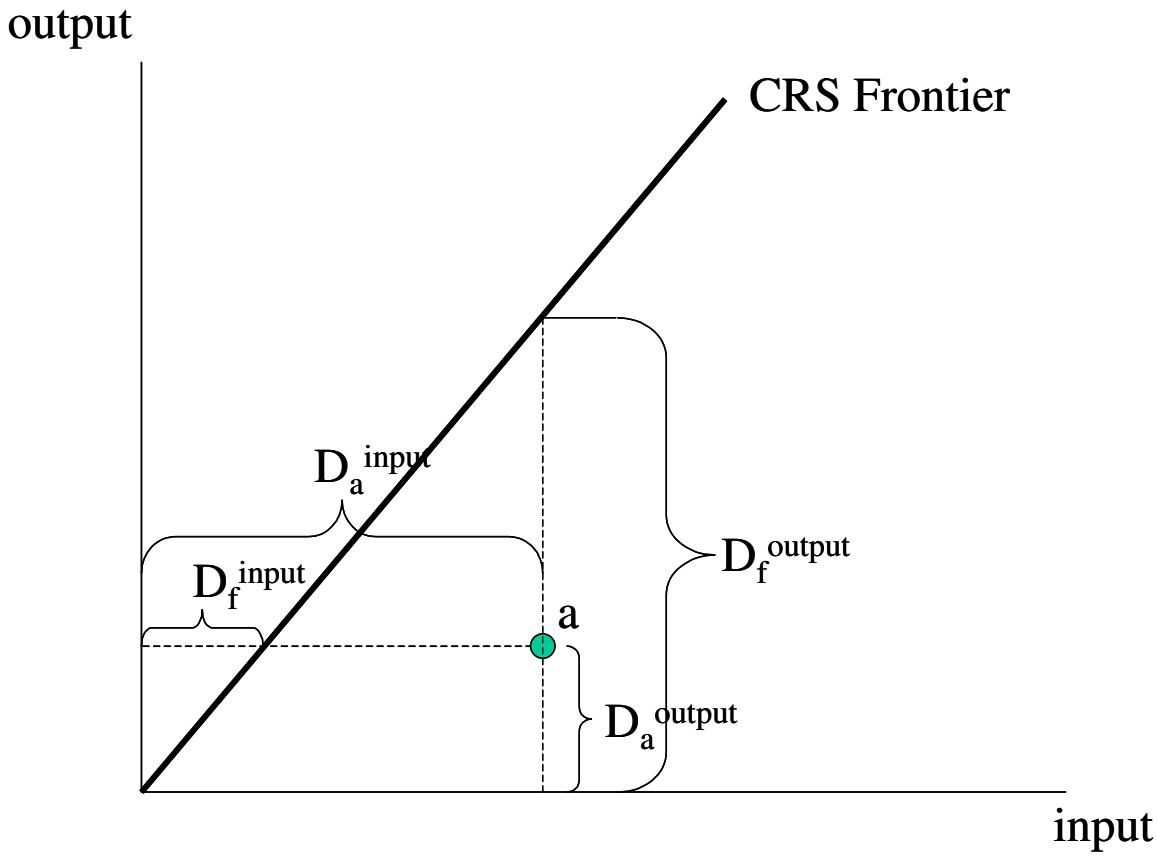

Figure 2.1: Distances used to calculate efficiency relative to a CRS frontier

The constant returns to scale model is appropriate when it is assumed all DMUs are believed to be operating at optimal scale. If this is not true then technical inefficiency can be split into a scale inefficiency effect and a technical inefficiency. The DEA model to measure technical inefficiency under variable returns to scale (VRS) is

$$
\begin{aligned}
&\min_{\theta,\lambda}(\theta), \\
&s.t. \quad -y_i + Y\lambda \geq 0, \\
&\qquad \theta x_i - X\lambda \geq 0, \\
&\qquad N1'\lambda \geq 0, \\
&\qquad \lambda \geq 0
\end{aligned}
\tag{2.13}
$$

where $N1$ is an $N \times 1$ vector of ones. The additional constraint limits the feasible region causing a tighter envelopment of the data to be constructed. Because now the frontier is at least as close to each data point as the previous CRS frontier the efficiency estimates generated are greater than or equal to the efficiency estimates calculated by CRS-DEA model. The VRS model has been stated for the input oriented model; however, the output oriented VRS model can be constructed by adding the convexity constraint to the output oriented CRS model.

Looking at the example shown in figure 2.2 we can see first the VRS frontier is strictly contained within the cone of the CRS frontier. Also for point e, evaluated from an input orientation, we can see the CRS frontier is constructed through a convex combination of point c and the origin. However, when evaluated against the VRS frontier point e is compared to a convex combination of point a and point b. Typically this additional constraint, in an input oriented model, causes convex combinations of DMUs more similar in output size to the DMU under consideration to be used as a frontier reference set.

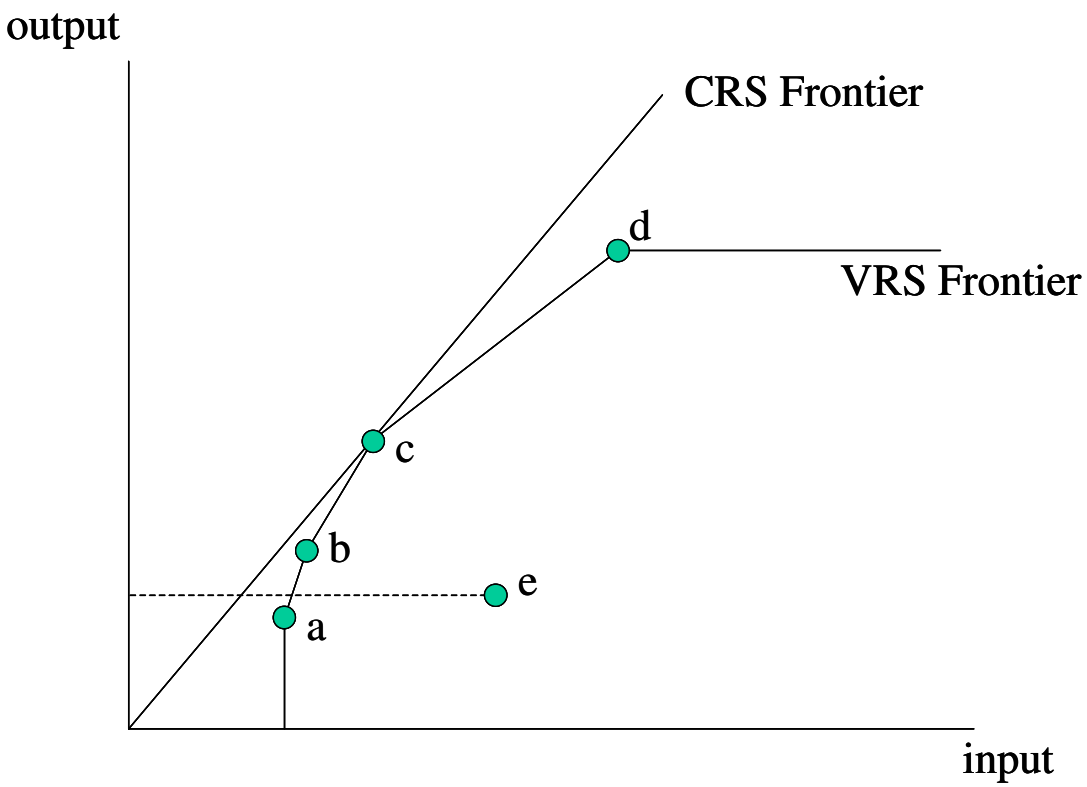

Figure 2.2: CRS frontier and VRS frontier shown for a single output and single input

By calculating both the VRS and the CRS to scale frontiers the scale inefficiency can be calculated, Banker [1984]. Efficiency of scale is equal to

$$\frac{\text{VRS Efficiency Score}}{\text{CRS Efficiecy Score}} \tag{2.14}$$

Notice for inefficient DMUs at a given level of output the value of scale efficiency is defined regardless of the input level of the point being evaluated.

Referring back to the postulates of DEA, the monotonicity postulate implies

(a) If $(x, y) \in T$ and $x_1 \geq x$, then $(x_1, y) \in T$.

(b) If $(x, y) \in T$ and $y_1 \leq y$, then $(x, y_1) \in T$.

This is another way of stating the concept of free disposability of outputs and inputs. In figure 2.3 the two dimensional section is shown. A frontier is constructed on which the output level for all DMUs is the same. Free disposability can be seen in figure 2.3 as the thick lines that construct part of the frontier and are parallel to the axis. This assumption implies that if you had more input of type one or two, it could be thrown away without cost. However, when efficiency is calculated, particularly efficiency over time, for DMUs that are compared to points on the surfaces created by free disposability the resulting measures are often distorted. The main argument is based the difference between the Koopmans [1951] and Farrell [1957] definitions of efficiency. Farrell would consider a' an efficient point, whereas, Koopmans would say it is not efficient. It seems natural to question the efficiency of a' because DMU a can produce the same amount of output using less of input 1.

Figure 2.3: Efficient frontier in input space with cost ratio planes identified

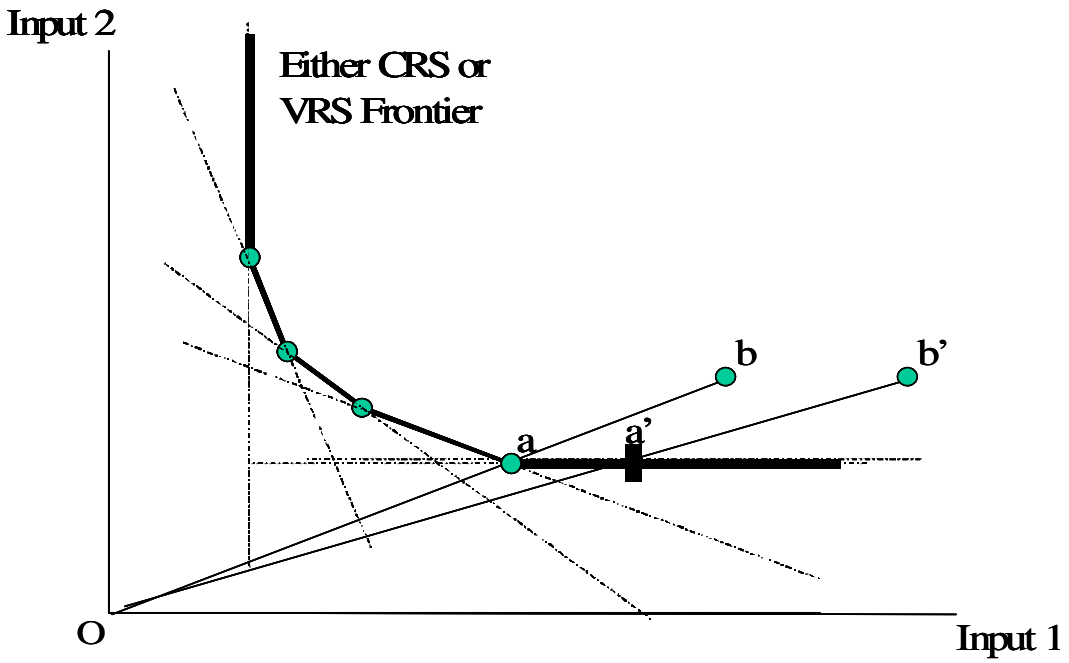

Figure 2.3: Efficient frontier in input space with cost ratio planes identified

Thus at point b', the value of reducing input 2 by one unit would be positive (in terms of efficiency improvement) and the value of reducing input 1 by a unit would not improve efficiency. For b' a reduction in input 1 to point b (while maintaining the same level of input 2) will be given a zero weight in an analysis of efficiency using DEA. This means while it is natural to think an observation would become more efficient by reducing input, in this case when b reduces input 1 it does not affect the DEA efficiency estimate. This can be seen by comparing $\frac{oa}{ab}$ to $\frac{oa'}{ob'}$. As long as b and b' efficiencies are measured relative to the portion of the frontier constructed by the assumption of free disposability, $\frac{oa}{ab} = \frac{oa'}{ob'}$. However, this concern is particularly troubling in a DEA time series analysis.

A related problem is the imputed marginal prices of resources in the free disposability portion of the production frontier. Note that the boundary is piecewise linear, and each linear segment represents a cost ratio between the two inputs. While DEA does not need to use cost data directly to calculate technical inefficiency, each one of these planes represents an implicit cost ratio. While the portions of the frontier for which the slope (i.e. marginal prices) is zero or ∞+ are the most egregious offenders, places with very shallow slopes or very steep slopes could also be considered unreasonable when compared to actual market prices.

2.2.1 Super Efficiency

In developing frontier proximity measures, a useful concept is super efficiency. Super efficiency estimates were first introduced by Anderson and Petersen [1993]. The super efficiency model measures the efficiency of an observation relative to a production possibility set constructed from a set of observations which does not include the observation under analysis. Typically DEA input oriented efficiency estimates range from zero to one; however, when the observation under analysis is not used in estimating the production possibility set, this is no longer the case. For observations outside of the production possibility set, efficiency estimates may be larger than one. Super efficiencies estimates can be computed using the DEA linear program shown below

$$\begin{aligned} &\min_{\theta_{jI},\lambda}(\theta_{jI}), \\ &s.t. \quad -y_j + \sum_{\substack{k=1 \\ k\neq j}}^{n} Y_k \lambda_k \geq 0, \\ &\qquad \theta_{jI} x_j - \sum_{\substack{k=1 \\ k\neq j}}^{n} X_k \lambda_k \geq 0, \\ &\qquad \lambda_k \geq 0 \text{ for all k} \end{aligned} \tag{2.15}$$

where DMU j is under evaluation, x_j is its P dimensional input vector and y_j is its Q dimensional output vector. θ_{jI} is a scalar defining the proportional increase or decrease of the j[th] DMU's input vector which is required in order to produce the j[th] DMU's output vector efficiently as defined by a frontier constructed using the n-1 other members of the reference set. λ is an intensity vector in which λ_k denotes the intensity of the k[th] unit. Here θ_{jI} is a super efficiency measure. This shows super efficiency estimates calculated for an input orientation, however, super efficiency estimates can be calculated for an output orientation also.

2.3 Exogenous Variable Models

There are many factors not under the control of the decision maker that affect efficiency. These are called exogenous or environmental variables. Some examples could be weather, market conditions, and other companies' competitive behaviors. These variables are exogenous for both the short-run and the long-run problem. However, there are decisions that can be made in long run planning that, once made, become exogenous variables to the short-run problem such as space or investment decisions. There are several models developed to handle these types of variables and they will be discussed below.

The first model to include attributes will be a social efficiency or "efficiency in use" model. In this model exogenous inputs are treated the same as endogenous variables.

$$\begin{aligned} &\min_{\theta_I,\lambda}(\theta_I), \\ &s.t.\ \ -y_i + Y\lambda \geq 0, \\ &\qquad \theta_I x_i - X\lambda \geq 0, \\ &\qquad \lambda \geq 0 \end{aligned} \tag{2.16}$$

X is a $(P+R)\times N$ matrix, where P is the number of endogenous input variables and R is the number of exogenous input variables. This model answers the question, "Considering as inputs the previous decisions of units or the conditions under which they operate, which unit is the most efficient at producing output?"

The two stage semi-parametric model suggested by Ray [1991] is a method to measure the effect of environmental variables on efficiency estimates. In the first stage, efficiency estimates (ES) are calculated using either CRS or VRS-DEA models using endogenous inputs and outputs in the formulation. The estimates generated in the first stage are then regressed against several environmental variables in the second stage using the following regression equation.

$$ES = \alpha + \beta_1 z_1 + ... + \beta_R z_R + \varepsilon \tag{2.17}$$

Here $(z_1, z_2, ..., z_R)$ are the environmental variables and the parameters $\alpha, \beta_1, ..., \beta_R$ are estimated. The maximum positive error term (ε) is then added to the intercept α resulting in α'. Adjusted efficiency scores (NES) are calculated as shown in (2.18):

$$NES = ES - \hat{\alpha}' - \hat{\beta}_1 z_1 - ... - \hat{\beta}_R z_R - \varepsilon \tag{2.18}$$

All values greater than one are truncated at one. The effect of this adjustment is to maintain the traditional understanding of efficiency as a value ranging from zero to one. The reason some adjusted values might be greater than 1 is those observations are performing well even though they are facing severe environments as measured by the environmental variables. The correction for this severe environment actually over corrects and causing some efficiency estimates to be greater than one.

Recently a bootstrapping technique to replace the second-stage regression has been introduced in Simar and Wilson [2005]. The need for a new technique was stated as two fold, 1) the current method lacks a coherent data-generating process and 2) mishandling of the complicated unknown serial correlation among the estimated efficiencies and the correlation between the ε_i and the z_i. The proposed methods uses the relationship between Shephard's input distance function and input efficiency. Shephard's input distance function is

$$\delta(x,y) = \left(\theta\left(x,y\right)\right)^{-1} = \sup\left\{\delta \middle| \frac{x}{\delta} \in X(y)\right\} \tag{2.19}$$

The input distance function δ gives a normalized measure of the distance from a point $\left(x,y\right)$ to the frontier, holding output and the direction of the input vector fixed. The following algorithm for the second-stage was suggested:

[1] Using the original data, compute $\hat{\delta}_i = \hat{\delta}\left(x_i y_i \middle| \hat{P}\right) \forall i = 1,...,n$ from (2.13) and (2.19).

[2] Use the method of maximum likelihood to obtain an estimate $\hat{\beta}$ of β as well as an estimate of $\hat{\sigma}_\varepsilon$ of σ_ε in the truncated regression of $\hat{\delta}_i$ on z_i in $\hat{\delta}_i = z_i\beta + \varepsilon_i \geq 1$.

[3] Loop over the next four steps ([3.1]-[3.4]) L_1 times to obtain n sets of bootstrap estimates $B_i = \left\{\hat{\delta}_{ib}^*\right\}_{b=1}^{L_1}$:

[3.1] For each i = 1,…,n draw ε_i from the $N(0,\hat{\sigma}_\varepsilon^2)$ distribution with left truncation at $\left(1 - z_i\hat{\beta}\right)$.

[3.2] Again for each i = 1,…,n compute $\delta_i^* = z_i\hat{\beta} + \varepsilon_i$.

[3.3] Set $x_i^* = x_i, y_i^* = y_i \frac{\hat{\delta}_i}{\delta_i^*}$ for all i = 1,…,n.

[3.4] Compute $\hat{\delta}_i^* = \hat{\delta}\left(x_i y_i \middle| \hat{P}^*\right) \forall i = 1,...,n$ where $\hat{P}^*$ is obtained by replacing Y, X in (2.13) and then using equation (2.19).

[4] For each i = 1,…,n compute the bias-corrected estimator $\hat{\hat{\delta}}_i$ defined by $\hat{\hat{\delta}}_i = \hat{\delta}_i - \text{BIAS}\left(\hat{\delta}_i\right)$ where

$\text{BIAS}\left[\hat{\delta}(x,y)\right] = B^{-1}\sum_{b=1}^{B}\hat{\delta}_b^*(x,y) - \hat{\delta}(x,y)$ were $\hat{\delta}_b^*(x,y)$ are the bootstrap estimates from step [3.4] and $\hat{\delta}_i$ is the original estimate.

[5] Use the method of maximum likelihood to estimate the truncated regression of $\hat{\hat{\delta}}_i$ on z_i, yielding estimates $\left(\hat{\hat{\beta}}, \hat{\hat{\sigma}}\right)$.

[6] Loop over the next three steps ([6.1]-[6.3]) L_2 times to obtain a set of bootstrap estimates $C = \left\{\left(\hat{\beta}^*, \hat{\sigma}_\varepsilon^*\right)_b\right\}_{b=1}^{L_2}$:

[6.1] For each i = 1,…,n draw ε_i from the $N\left(0, \hat{\hat{\sigma}}\right)$ distribution with left-truncation at $\left(1 - z_i\hat{\hat{\beta}}\right)$.

[6.2] Again for each i = 1,…,n compute $\delta_i^{**} = z_i\hat{\hat{\beta}} + \varepsilon_i$.

[6.3] Use the maximum likelihood method to estimate the truncated regression of δ_i^{**} on z_i, yielding estimates $\left(\hat{\hat{\beta}}^*, \hat{\hat{\sigma}}^*\right)$.

[7] Use the bootstrap values in C and the original estimates $\hat{\hat{\beta}}, \hat{\hat{\sigma}}$ to construct estimated confidence intervals for each element of β and for σ_ε.

The confidence intervals can be constructed for any of the parameters including the efficiency estimate in the following manner. To find a confidence interval for a particular β_j consider the value $\left(\hat{\hat{\beta}}_j - \beta_j\right)$. A confidence interval for β_j can be defined as

$$\Pr\left[-b_\alpha \le \left(\hat{\hat{\beta}}_j - \beta_j\right) \le -a_\alpha\right] = 1-\alpha \tag{2.20}$$

If the distribution of β_j were known then it would be straightforward to find a_α, b_α, however, it is not, so the bootstrap assumption

$$\left(\hat{\hat{\beta}}_j - \beta_j\right) \Big| P \overset{approx}{\sim} \left(\hat{\hat{\beta}}_j^* - \hat{\hat{\beta}}_j\right) \Big| \hat{P} \tag{2.21}$$

is used. Thus equation (2.15) is approximated by

$$\Pr\left[-b_\alpha^* \le \left(\hat{\hat{\beta}}_j^* - \hat{\hat{\beta}}_j\right) \le -a_\alpha^*\right] \approx 1-\alpha \tag{2.22}$$

This approximation improves in a statistical sense as the number of iterations in the bootstrap, L_2, increases.

The new method improves the original two-stage model that only allowed point estimations of efficiency. By constructing confidence intervals, the uncertainty related to efficiency estimates based on an estimate of an unobservable frontier can begin to be quantified. A confidence interval provides the analyst significantly more information than the point estimate.

Alternatively, the Lovell-Ruggiero model can also be used to calculate efficiency scores in the presence of environmental or exogenous variables. The linear program used by Ruggiero to calculate efficiency scores is

$$\begin{aligned}
&\min_{\theta_I,\lambda}(\theta_I),\\
&s.t.\ \ -y_i + Y\lambda \geq 0,\\
&\quad \theta_I x_i - X\lambda \geq 0,\\
&\quad \text{if } z_l^j > z_l^i \text{ then } \lambda^j = 0,\\
&\text{where } l \text{ is an index of the exogeneous variables}\\
&\quad N1'\lambda = 1,\\
&\quad \lambda \geq 0
\end{aligned} \tag{2.23}$$

This alternative incorporates the idea of creating appropriate comparison groups. Here only DMUs in equal or worse conditions for all environmental factors are allowed to be members of the comparison group for a given DMU.

A third alternative is a three-stage model described in Fried et al. [2002]. In this model DEA efficiency estimates are generated in the first stage using CRS or VRS. Either an input or an output orientation can be taken. In the second stage the total slacks in the input and output constraints, $[x - X\lambda] \geq 0$ and $[Y\lambda - y] \geq 0$ are considered. These slacks are interpreted as being composed of three effects: environmental influences, managerial inefficiencies, and statistical noise or measurement error. SFA is then used to estimate values for these components. An SFA regression is run for each of the N input

constraints $s_{ni} = x_{ni} - X_n\lambda \geq 0,\ n = 1,...,N,\ i = 1,...,I$ where s_{ni} is the stage 1 slack in the usage of the n[th] input for the i[th] producer. The independent variables for the SFA regression model are the elements of the R observable environmental variables, $z_i = [z_{1i},...,z_{Ri}],\ i = 1,...,I$. The N separate Stage 2 SFA regressions are

$$s_{ni} = f^n(z_i;\beta^n) + v_{ni} + u_{ni},\ n = 1,...,N,\ i = 1,...,I \tag{2.24}$$

Assume $v_{ni} \sim N(0,\sigma_{vn}^2)$ reflects statistical noise and $u_{ni} \sim N^+(\mu^n,\sigma_{un}^2)$ reflects managerial inefficiency. All parameters $(\beta^n,\mu^n,\sigma_{vn}^2,\sigma_{un}^2)$ are allowed to vary across the N input slack regressions. From these results adjusted inputs can be calculated as

$$x_{ni}^A = x_{ni} + [\max_i\{z_i\hat{\beta}^n\} - z_i\hat{\beta}^n] + [\max_i\{\hat{v}_{ni}\} - \hat{v}_{ni}],\ n = 1,...,N,\ i = 1,...,I \tag{2.25}$$

where x_{ni}^A and x_{ni} are adjusted and observed input quantities, respectively. Before this calculation can be done v_{ni} must be determined. From the conditional estimators for managerial inefficiency given by $\hat{E}[u_{ni} \mid v_{ni} + u_{ni}]$, estimators for statistical noise are derived residually by means of

$$\hat{E}[v_{ni} \mid v_{ni} + u_{ni}] = s_{ni} - z_i\hat{\beta}^n - \hat{E}[u_{ni} \mid v_{ni} + u_{ni}],\ \ n = 1,...,N,\ i = 1,...,I \tag{2.26}$$

which provide conditional estimators for the v_{ni}. In the third stage the adjusted inputs and original outputs are used in the same DEA model as used in the first stage. The results of the model are managerial efficiency scores, without the effects of operating environment and statistical noise.

2.4 Defining time series methods

DEA was initially a method for measuring efficiency within a cross section of data. However, it is often interesting not only to compare data across groups of firms but also

compare efficiency changes over time. Cross sectional data available at several points in time, is called *panel data*. Two indices are associated with each observation: k to indicate the unit, and t to denote the point in time when the observation is made. An observation is stated $\left(x_{kt}, y_{kt}\right)$ and the data set is described as

$$Y_{KT} = \{(x_{kt}, y_{kt}) \mid x_{kt} \in R_+^I, y_{kt} \in R_+^J;\ \text{k=1,2,...,n;}\ \text{t=1,2,...,m}\} \tag{2.27}$$

where the subscript K and T refer to the sets of firms and of observation times.

Typically when efficiency measurement is performed, a DMU is compared to all DMUs operating during the same time period. Here the time period defines the *reference set*, the set of data used for comparison purposes when measuring efficiency. When evaluating efficiency in panel data, there are several ways to define the reference set, and the terminology outlined in Tulkens and Vanden Eeckaut [1995] will be followed. The first is to construct a reference set at each point in time, t, from the observations made at that time only. The reference set

$$Y_{Kt} = \{(x_{kt}, y_{kt}) \mid \text{k=1,2,...,n}\} \tag{2.28}$$

is used at each point in time t=1,2,...,m. This is called *contemporaneous* and the production set is stated as

$$Y\left(Y_{Kt}\right),\ \text{t=1,2,...,m} \tag{2.29}$$

A reference set also can be constructed at each point in time t, using the observations made at times $s=1$ up until $s=t$. The reference set at each time $t=1,2,...,m$ is

$$Y_{K(1,t)} = \{(x_{ks}, y_{ks}) \mid x_{kt} \in R_+^I, y_{kt} \in R_+^J;\ \text{k=1,2,...,n;}\ s\text{=1,2,...,t}\} \tag{2.30}$$

m successive reference sets can be defined in this way and will be called *sequential*. The corresponding production set is stated as

$$Y\left(Y_{K(1,t)}\right),\ \text{t=1,2,...,m} \tag{2.31}$$

A third possible construction of the reference set is to include observations made throughout the whole observation period. The reference set is Y_{KT} and the production set, which will be called *intertemporal*, is $Y\left(Y_{KT}\right)$.

The most popular method of using DEA efficiency scores for doing comparisons over time is the Malmquist index. Introduced by Fare et al. [1994] as a method to compare two time periods t and t+1, Malmquist allows estimates of technical inefficiency and technical progress. Taking time period t as the reference period, the input oriented Malmquist index (IM) is

$$IM^{t} = \left[\frac{D_i^t(y^{t+1}, x^{t+1})}{D_i^t(y^t, x^t)}\right] \tag{2.32}$$

$D_i^t(y^t, x^t)$ defines the technology of production in terms of an input distance function , as follows

$$D_i^t(y^t, x^t) = \sup_{\delta}\{\delta : \left(x^t \big/ \delta\right) \in L^t(y^t), \delta > 0\} \tag{2.33}$$

where the subscript i denotes input orientation. $D_i^t(y^t, x^t)$ is the reciprocal to Farrell's input oriented measure of technical efficiency. IM^t compares (y^{t+1}, x^{t+1}) to (y^t, x^t) by measuring their respective distances from the constant returns to scale (CRS) production boundary of the reference period t. In a similar fashion, with reference to period t + 1, one may define the following index:

$$IM^{t+1} = \left[\frac{D_i^{t+1}(y^{t+1}, x^{t+1})}{D_i^{t+1}(y^t, x^t)}\right] \tag{2.34}$$

IM^{t+1} measures the distance of (y^{t+1}, x^{t+1}) and (y^t, x^t) from the CRS production boundary of period t + 1. To avoid an arbitrary choice of a reference period, Fare et al. [1994] uses the geometric mean of IM^t and IM^{t+1} resulting in

$$IM = \left[\frac{D_i^t(y^{t+1}, x^{t+1})}{D_i^t(y^t, x^t)} \frac{D_i^{t+1}(y^{t+1}, x^{t+1})}{D_i^{t+1}(y^t, x^t)} \right]^{1/2} \tag{2.35}$$

This can be factored to show

$$IM = \frac{D_i^t(y^{t+1}, x^{t+1})}{D_i^t(y^t, x^t)} \times \left[\frac{D_i^t(y^{t+1}, x^{t+1})}{D_i^{t+1}(y^{t+1}, x^{t+1})} \frac{D_i^t(y^t, x^t)}{D_i^{t+1}(y^t, x^t)} \right]^{1/2} = \Delta TE \times \Delta TC \tag{2.36}$$

where ΔTE measures technical efficiency change and ΔTC measures the geometric mean of the magnitude of technical change. Because of the relationship between distance functions and Farrell's technical efficiency measure, each one of these distance functions can be calculated by solving a DEA linear program. Other methods have been used to estimate the distance function, so for clarity this method will be called input oriented DEA Malmquist.

In the following chapters the methods described here will be applied. In Chapter 4, the super efficiency model and the two-stage model are applied to develop an outlier detection method. In the second stage of the two stage model the bootstrapping method is used. In Chapter 5 a variant of the super efficiency model is described. This new super efficiency model takes a non-traditional orientation to quantify super efficiency. The benefits of this model are the outlier detection technique and the Malmquist index measure can be calculated given all data are positive. This was not the case with the previous super efficiency models. In Chapter 6, Shephard's distance function is used to aggregate information about the inputs to create a regression model which can

characterize the multi-input / multi-output nature of production. This summarizes the basic results and where they are used in subsequent chapters.

CHAPTER 3

LITERATURE REVIEW

While the previous chapter introduced the mathematical tools and gave a brief description of the methods used in efficiency analysis, this chapter gives some historical background on the development of these techniques with particular emphasize on the literature related to exogenous variables. In the productivity measurement literature there is significant overlap between results presented by separate research groups with no reference by either to the work of the other group. In this way prior work is often not recognized properly and credit for developing particular methods not always properly placed. This chapter also describes some of these areas of the literature.

While this serves as a summary of the key literature in the productivity field a more extensive chapter length review of productivity and the methods of measurement, Lovell [1993] is recommended. If the reader would prefer a longer but not necessarily a rigorous mathematical treatment, Coelli, Rao and Battese [1998] is recommended. Their book contains three chapters explaining the background of production economics, and two chapters each on: DEA, SFA, and index numbers. The book ends with a chapter on the integration of productivity measurements and efficiency measurements. Both the chapter and the book recommended are written by economists and more easily understood if the reader has some background in economics.

3.1 Overview of Data Envelopment Analysis

A linear programming technique for determining efficiency, named Data Envelopment Analysis developed by Charnes, Cooper, and Rhodes, has roots in a field called activity analysis, which was a popular topic for econometricians in the 1950s. A key researcher in this field was Tjalling C. Koopmans. In Koopmans [1951] he developed his definition of efficiency.

> "A possible point in the commodity space is called efficient whenever an increase in one of its coordinates (the net output of one good) can be achieved only at the cost of a decrease in some other coordinate (the net output of another good)."

Thus a technically efficient producer could produce the same output with less of at least one input, or could use the same inputs to produce more of at least one output. He also introduced the idea of a piecewise linear frontier which could be defined by solving a set of linear equations. Michael J. Farrell later cited Koopmans' work in activity analysis as inspiring his ideas.

Working along side Koopmans at the Cowles Commission at the University of Chicago was Gerard Debreu. Debreu contributed to this field in Debreu [1951] which defined a coefficient of resource utilization as the ratio between minimized resource costs of obtaining a given consumption bundle and actual costs. This definition of efficiency in a cost context was generalized by Farrell to concept of production efficiency. Debreu also measured a proportional contraction of resources. Farrell also sited Debreu and his coefficient as an inspiration for the concept of technical efficiency; however, it should be noted Debreu's concept was built strictly from resource cost side analysis.

The path breaking paper which introduced the field of non-parametric efficiency was Farrell [1957]. The contributions of this paper can be summarized as three fold:

(1) an efficiency measure based on radial contractions or expansions from inefficient observations to the frontier;

(2) a production frontier specified as the most pessimistic piecewise linear envelopment of the data; and

(3) the newly defined frontier calculated through solving systems of linear equations, obeying the conditions on the unit isoquant that its slope is not positive and that no observed point lies between it and the origin.

His efficiency measure could be further divided into technical efficiency and price (or allocative) efficiency. It should be noted that Farrell's definition of technical efficiency is weaker than Koopmans'. This has lead to many articles discussing this issue see, e.g. Lovell [1993]. Farrell also begins to articulate the concept that would later be known as the duality between cost and production functions when he noted that his efficiency measure also had a cost interpretation. The duality concept was also describe in Shephard [1953] and Shephard [1970] where Shephard's distance function was used as an intermediate step in showing this relationship mathematically.

In the Discussion of Farrell's paper, A. J. Hoffman observes that to completely describe the frontier defined by Farrell's method "m (number of inputs) and n (number of outputs) do not have to be very large for the problem to become hopeless." However, he noted if the goal is rather to calculate efficiency for a given point this could be formulated as a linear program and solved by C.E. Lemke's newly developed dual simplex method. It is interesting to note that Lemke was the first Ph.D. student of Abraham Charnes who is the first author of the seminal paper Charnes, Cooper and Rhodes [1978] (CCR). Farrell took Hoffman's suggestion and in Farrell and Fieldhouse

[1962] developed a constant returns to scale model using the unit isoquant. Even though they did not implement the multi-output case the generalization of the constant returns to scale model was described.

Contrary to popular belief the multiple input and multiple output linear programming efficiency model is not due to CCR. Rather it was develop nearly ten years earlier by an agricultural economist at the University of California at Berkeley named James N. Boles. In Boles [1966], a paper he presented at the Western Farm Economics Association annual meeting held in Los Angles, California is found the single-output Farrell model and the correct ideas for the generalization. It is interesting to note that Boles does not cite either Hoffman's comment on Farrell's paper or Farrell and Fieldhouse, when it appears he most definitely knew of this work as noted by Forsund and Sarafoglou [2002]. Here begins what seems to be a tradition in this area of research, namely, of failing to cite appropriate precedents in the known literature. Boles also presented the dual to the linear program for calculating efficiency estimates and gave the cost interpretation:

> "An economic interpretation of the dual is to select a set of nonnegative factor prices to minimize the cost of producing one unit of the jth activity subject to the condition that the cost of production of each of the n activities is greater than or equal to 1.0. For those activities in the optimal basis, the cost of production will be equal to 1.0, and the efficient facet coincides with a unit isocost hyperplane with all points, P_k, on the hyperplane or on the side away from the origin k=1,2,…,n."

Later, Boles [1971] showed computer programs to handle three types of problems: "single product with no economies or diseconomies of scale, multiple product with no economies or diseconomies of scale, and single product with economies of scale." In this paper, a model with multiple outputs and inputs identical to the model termed "ordinary linear programming problem" by CCR can be found. Boles includes his complete

computer codes in FORTRAN, which not only output efficiency estimates, but also slacks and shadow prices on the constraints.

William Cooper working with his Ph.D. student Ewardo Rhodes was investigating ratio form or fractional programming problems to evaluate the efficiency of a public schools program. While investigating methods for efficiency evaluation Rhodes discovered Farrell's 1957 paper and the two began to adapt their fractional programming methods to create the linear programming technique call data envelopment analysis (DEA). While the term DEA appears in Rhodes [1978] dissertation interestingly it only appears in the classical reference Charnes, Cooper and Rhodes [1978] when referring to Rhodes' dissertation.

While much of the credit given to Charnes, Cooper and Rhodes (1978) appears inappropriate, in light of Boles' overlooked contribution, it is valuable to list the many contributions of CCR that are not so often stated. CCR identified the connection between productivity indices, which tried to find weighted sums of inputs and outputs, and the Farrell's technical efficiency measure. While Farrell's description of his technique was simple and intended to be easily understood by all, as stated in his paper, CCR added the mathematical rigor allowing DEA to be the valuable research tool it is today. Finally they gave the explicit interpretation of the primal and dual problems including the economic meaning of shadow prices. The model developed by CCR assumed constant returns to scale (CRS) so this model will be referred to as the constant returns to scale DEA (CRS-DEA) model.

While the research on linear programming methods to calculate efficiency estimates between the time of Farrell and CCR is often overlooked, this literature review attempts

to give a more complete summary. For a detail account of the history of DEA, please see Forsund and Sarafoglou [2002].

3.2 Variable returns to scale model

While the variable returns to scale (VRS) model is typically attributed to Banker, Charnes and Cooper [1984] or Banker [1984], here again is another instance of unrecognized prior results. A discussion of variable returns to scale should begin, or at least include, the work of Sidney N. Afriat. In Afriat [1972] while exploring efficiency estimates of production functions, Afriat identifies the exact constraint, the convexity constraint, which distinguishes Banker's VRS model from CCR's constant return to scale (CRS) model.

$$\sum \lambda_i = 1 \tag{3.1}$$

While Afriat's initial work was in demand analysis, the 1972 paper develops "a function that gives maximum feasible output, so observed output must be bounded by it from above, in a ratio, which measures efficiency". This sounds very similar to minimum convex hull that envelops the data, discussed by CCR; however, Afriat in his exploration of constrained optimization models never suggested the CCR model.

3.3 Parametric models

One of the foundations of the parametric method is corrected (ordinary) least squares (COLS). To begin, *linear regression* is a method of identifying a linear function on a set of data that characterizes the relationship between the attributes of the data points for the population. *Ordinary least squares* is a particular linear regression method in which the objective is to minimize the sum of squared residuals. Call y a dependent variable and

x_k a vector of independent variable. Let there be n observations of y and x_k which are indexed by i. X is then the $n \times K$ matrix of independent variables. The regression model is $E\left[y_i \middle| x_i \right] = x_i' \beta$, and an estimate of $E\left[y_i \middle| x_i \right]$ is denoted

$$\hat{y}_i = x_i' b \tag{3.2}$$

The *disturbance* associated with the ith data point is

$$\varepsilon_i = \hat{y}_i - x_i' \beta \tag{3.3}$$

For any value of b, an estimate of ε_i called the *residual* can be calculated

$$e_i = \hat{y}_i - x_i' b \tag{3.4}$$

The ordinary least squares parameter estimation is computed from the optimization problem

$$\underset{b}{Minimize}\ S(b) = e'e = (y - Xb)'(y - Xb) \tag{3.5}$$

Where y is now a $n \times 1$ matrix of dependent variables. X can be augmented by a column of 1s so that now X is $n \times (K+1)$. Now the minimization problem outputs a $1 \times (K+1)$ vector b, called the *coefficients*, where the first term b_0 is associated with the vector of ones.

A deterministic frontier can be consistently estimated simply by increasing the constant term until the largest residual goes to zero Gabrielsen [1975]. The corrected constant term is $b_0 + \max_i(e_i)$. The resulting efficiency measures are

$$\hat{\theta}_i = e_i - \max_i(e_i) \tag{3.6}$$

Another parametric approach suggested by Aigner and Chu [1968] falls within deterministic frontier models. This model uses linear programming to minimize the sum of strictly negative error terms. The formulation is

$$\begin{aligned} &\min_{\beta} \sum_{i} |\varepsilon_i| \\ &s.t. \quad \varepsilon_i = y_i - \alpha - \sum_{k} \beta_k x_{ki} \le 0 \ \forall i \end{aligned} \tag{3.7}$$

The main drawback of this model is the estimates (β_i) are not consistent in a statistical sense and hence the residuals or the inefficiency estimates (ε_i) are also biased. It may be argued that by not specifying a distribution for ε_i these estimates could be robust to a variety of distributional assumptions, though this remains to be verified. While for a time these models were popular, the concern for measurement error or chance occurrences suggested a different model was needed.

Stochastic frontier approach (SFA) was developed to allow for random error, so the purpose of these models was not only to measure inefficiency but also to identify random error. Neither random error nor inefficiency can be observed, so separating them requires an assumption. SFA employs a *composed error* model in which inefficiencies are assumed to follow an asymmetric distribution, usually the half-normal, while random errors are assumed to follow a symmetric distribution, usually the standard normal, see Aigner, Lovell and Schmidt [1977]. That is, the error term is given by $\varepsilon = \mu + \nu$, where $\mu \ge 0$ represents inefficiency and follows a half-normal distribution, and ν represents random error and behaves according to a normal distribution. Jondrow et al. [1982] developed a means of finding the individual efficiencies for each data point. Using the conditional distribution of the asymmetric error given the whole error, the two

components can be separated. While the shape of the distribution of the inefficiency term has been the focus of research and debate, SFA has the desirable property that regardless of the distribution imposed, the ordering of the observations does not change significantly.

Individual specific effects (ISE) as a part of regression analysis have been observed in the econometric literature since at least Mundlak [1961] where the decomposition of the error and estimation via fixed effects and within estimators were demonstrated. This concept of individual specific effects can be adjusted and redefined to be a firm specific inefficiency term as is done in Schmidt and Sickles [1984]. To apply this method, panel data is used and a functional form for the production function is specified. A common choice is the translog function. In the model, the level of inefficiency is assumed constant over the time period and inefficiency is the only time-invariant fixed effect. While these assumptions may be considered drawbacks, some benefits are that the distribution of efficiency over all firms is not prespecified as in SFA and that in a statistical sense the technical inefficiency can be estimated consistently. While there are many estimation methods for this type of model, here only the within model will be discussed. Efficiency is estimated using the deviation from the most efficient firm's intercept term. However, as inefficiency is no longer a separately specified element in a composed error term, the assumption that inefficiency is uncorrelated with the regressors (as in SFA) is no longer needed. As larger data sets have become more easily accessible in recent years, this method has grown in popularity.

Thick frontier approach, first developed by Berger and Humphrey [1991], starts by sorting the data based on average costs, where average cost is total cost divided by total

assets. Then two "thick-frontiers" are estimated, one for the lowest and one for the highest cost quartiles of DMUs. These regressions are independently executed. Deviations from predicted performance values within the highest and lowest performance quartiles of firms represent only random error, while deviations in predicted performance between the highest and the lowest average-cost quartiles represent only inefficiencies plus exogenous differences in the regressors. A separate efficiency score for every DMU can be calculated. The estimated residuals for the entire sample are calculated and the inefficiency disturbances are assumed to be uncorrelated with the regressors, so that a separate intercept for each DMU can be recovered as the mean of its residuals. This method depends heavily on arbitrary assumptions, such as the variation within the lowest cost quartile is attributed entirely random error and the difference between the lowest quartile and the upper quartile is attributed to inefficiency.

3.4 Methods for handling exogenous variables

Hall and Winsten [1959] were the first to recognize and name environmental variables in a Farrell type model, demonstrating the relationship of such variables to various types of efficiency. They identified social efficiency or "efficiency in use" for which environmental variables are treated as any other input or output variable. Then they considered managerial or technical efficiency for which ordering of observations through productivity measures can only be done for each environmental setting. Then the problem of merging scales arises when a comparison is done across environments. The authors also note that the specification of environmental variables may be different when doing long-run evaluations or short-run evaluations. While convexity is an assumption of Farrell's model, Hall and Winsten argue against this assumption in some cases. Hall and

Winsten also discern the difference between explanatory analysis (regression) and efficiency analysis. Linear programming efficiency techniques assume there is a set of methods of production, each of which has constant technical coefficients, i.e., a given amount of each input per unit of output. While technical inefficiency arises from combining these methods of production in the wrong proportion, managerial efficiency needs a measure of hardness or difficulty of the task. To illustrate this point they use the example of forecasting. A coefficient intended to summarize the achievement of the forecaster is

$$f_1 = \frac{1}{T}\sum_{t=1}^{T}(v_t - r_t)^2 \tag{3.8}$$

Here v_t is the forecast for the year t, and r_t is the realized value for year t. A second coefficient is calculated

$$f_2 = \frac{1}{T}\sum_{t=1}^{T}(r_t - r_{t-1})^2 \tag{3.9}$$

intended to represent the difficulty of forecasting. Then an index

$$\Psi = \sqrt{\frac{f_1}{f_2}} \tag{3.10}$$

is calculated which measure the achievement relative to the difficulty of the task. While this index measure is not perfect (giving a high difficulty measure, f_2, to a uniformly increasing demand function), it demonstrates a situation in which a measure of difficulty was used to adjust a measure of quality. Similarly, Hall and Winsten argue managerial efficiency needs this type of scaling with consideration for the difficulty of the task.

Banker and Morey [1986b] introduced methods to handle categorical variables in DEA when they are either controllable or uncontrollable (this model will be called the

Banker and Morey (BM) categorical variable model). In the case of the controllable variable it is necessary to change the linear programming formulation of DEA to a mixed integer formulation. It is interesting to note that in Banker and Morey [1986a] the authors attempt to make DMUs in different environments comparable by decreasing the number of constraints in which the radial measure of contraction (typically θ is used) appears. Decreasing the number of constraints in which the radial measure appears causes efficiency estimates to stay the same or decrease. However, in Banker and Morey [1986b] the authors restrict comparison to DMUs in equal or worse conditions using a categorical uncontrollable variable, which means efficiency estimates either stay the same or increase. It would seem these two models are contradictory.

Banker and Morey [1986a] introduce a single stage method for handling environmental variables (this model will be called the Banker and Morey (BM) environmental variable model). This was the first method to incorporate exogenous variable values into a DEA calculation. Previously the method for handling exogenous variables was to run a DEA model for each group of DMUs at a given exogenous variable value level. However, as stated previously the BM environmental variable model does not change the shape of the production frontier, rather it simply limits the number of constraints in which the radial measure of contraction appears.

The single stage method can be contrasted with the two stage method that was first introduced by Timmer [1971]. While the main purpose of Timmer's paper was to introduce a probabilistic frontier production function for the purposes of measuring efficiency, in the three pages prior to the conclusion of his paper, he attempts to explain the efficiency estimate by regressing the estimates as dependent variables against a

variety of factors beyond simple inputs and outputs. Timmer notes the key flaw in the two-stage method "[the independent variable] makes this correction at the efficiency stage rather than at the more appropriate level, that is, where the input variables were constructed." Or stated differently, the two stage method attempts to explain efficiency estimates based on second stage regression variable; however, if the analyst thought the variable influenced the efficiency estimate, then why was the variable not included in the first stage? And since it was not, does the first stage calculation not suffer from omitted variable bias?

Ray [1988] introduces a two-stage model where in the first stage DEA is used to calculate the efficiency of a given unit and in the second stage the efficiency estimates are regressed against the environmental variables using a stochastic frontier (SFA) model. While the mathematical statement of his model is clear, when addressing the question of, why a variable should be used in the first stage or the second stage, Ray simply states "Only the discretionary inputs are to be included in the DEA stage. Unfortunately, there is no simple answer as to how one determines if a specific input is discretionary or not." Ray [1991] is a follow-up paper where he used a regression model rather than the SFA model (this model will be called Ray's two-stage model or two-stage model). In the second stage the efficiency estimates (ES) are regressed against the R different environmental variables. Here all z_i are oriented such that positive changes in z_i represent a more favorable environment, so that $\beta_i > 0$ would confirm a more favorable environment is correlated with higher efficiency. The index of efficiency estimates can be transformed by adjusting the intercept α so that NES is non-negative. The primary advantage of this method is the second stage allows for sensitivity analysis and different

sets of non-discretionary inputs can be tested. However, a drawback is the regression requires a priori specification of functional form. Also adjustments are made based on the two-sided error term, thus it is possible that inefficiency will be overstated. Further because the efficiency estimates are not calculated independently, rather they are all used to determine the frontier, the efficiency estimates are not independent and the residuals are not independent. It is interesting to note that while Ray properly cites Timmer's work in his 1988 paper, his much more widely known and read 1991 paper fails to mention Timmer's work.

The two-stage method has recently been criticized by Simar and Wilson [2005] for its lack of a coherent data-generating process and mishandling of the complicated unknown serial correlation among the estimated efficiencies. Their bootstrapping technique was described in section 2.3.

In McCarty and Yaisawarng [1993], the two stage model is compared to a single stage model. Their single stage model answers the question, "Given factors both within and beyond a DMU's control, how efficient is the DMU?" They answer this question using a DEA model, which treats controllable and uncontrollable inputs identically. They were interested in the effect of environmental variables in the case of New Jersey public schools. After the Abbott vs. Burke ruling identifying 28* schools as being under funded causing the students not to receive a "thorough and efficient" education, the authors wanted to investigate the claims of the defendant (the New Jersey school system). The defendant claimed that the plaintiff (4 of these 28 schools that felt they were under funded) schools were poorly managed and the allocation of additional funds was unlikely to improve matters much. The findings of this paper indicate that the two different methods produce rankings of efficiency estimates that are positively and significantly

correlated. The authors point out the purpose of the two models is different. Their discussion is very similar to Hall and Winston's discussion of social efficiency (efficiency in use) vs. managerial efficiency.

Lovell [1994] presented a model for handling uncontrollable inputs (very similar to Banker and Morey 1986a) by constraining the comparison set to units with the same or lower value for uncontrollable inputs. Lovell does not restrict his consideration to categorical variables, but rather considers any uncontrolled input. This is the same idea Ruggiero [1996] presents, however, Ruggiero follows an axiomatic approach, by which he arrives at the conclusion of allowing equal or worse environment units to create the reference set. This model will be referred to as the Lovell-Ruggiero exogenous variable model or Lovell-Ruggiero model.

Ruggiero [1996] notes that Banker and Morey [1986a] do not properly construct the production frontier when considering environmental variables. By allowing convex combinations of environments better and worse than a given environment, a DMU's efficiency may be evaluated based on an environmental state it could never realize. Thus Ruggiero suggests limiting the DMUs used to construct the frontier, allowing only DMUs with equal or worse environments than the DMU being evaluated. However, this severely limits the data used in the analysis of any given DMU.

Ruggiero recognized this shortcoming of his 1996 model and revisited the model in Ruggiero [1998], this time adjusting the set of DMUs used to construct a given frontier. In the case when there are multiple environmental variables, instead of limiting the comparison set to only DMUs with equal or worse environment variables in all environment variable dimensions, he suggests aggregating the environment variables of

each DMU into a single environment measure. This model will be referred to as Ruggiero's environment measurement model. Then a given unit would be compared to any unit with an equal or worse environmental measure. This measure is constructed by estimating efficiency based on the controllable inputs and outputs, regressing these efficiency estimates on the environmental variables in a multi-regression, calculating the regression coefficients and calculating the environment score by summing over all environmental variables the product of the regression coefficient times the environmental variable value for each unit. While this is a slight improvement compared to his initial model there is still a significant lack of data when calculating the efficiency of unit s with low environmental measures particularly if the environmental variables are continuous.

There are several possible problems that can arise with either the two-stage or the one-stage method. As McCarty and Yaisawarng [1993] warned, the two-stage approach could be problematic when there is strong correlation between the independent variables in the two stages and the claim that the second stage incorporates fundamentally different types of inputs, controllable and uncontrollable variables, becomes untenable. Ruggiero [2004] addresses a related problem considering the case when non-discretionary factors are correlated with technical efficiency. When the two are negatively correlated the efficiency estimates will be inflated. He suggests a modification to Ruggiero's environment measurement model to allow the comparison reference set to be expanded to account for this correlation.

3.5 Comparison of efficiency models

In Chapter 4 of this study, two stage models handling environmental variables will be demonstrated in a DEA setting. There are other efficiency measurement techniques, such

as stochastic frontier approach, free disposability hull approach, thick frontier approach, individual specific effects and others. In the academic research it is not clear any one method is preferred to another even given special conditions. The decision to use one method rather than another is largely ad-hoc and closely related to the analyst's personal preference. Some examples of comparisons from the literature will be given here to present the reader with a set of resources, which may help them to select which technique is proper for their needs.

Extensive comparative investigation has been recorded in the archival literature; however, the typical methodology is to assume a data generation process, generate efficiency estimates using two methods, then compare the results to the known efficiencies and draw conclusions about which method is better. The most widely cited comparison was performed by Gong and Sickles [1992] using panel data. Their results indicate that for simple underlying technologies, the relative performance of the stochastic frontier models vs. DEA depends on the choice of functional forms. If the employed form is close to the given underlying technology, stochastic frontier models outperform DEA using a number of metrics. As the misspecification of the functional form becomes more serious and as the degree of correlatedness of inefficiency with regressors increases, DEA's appeal becomes more compelling.

This method of generating data to evaluate the methods is flawed because defining the data generating process implicitly defines a functional form, which if known in the analysis, can be replicated by the stochastic frontier method to improve the results. Part of the argument for data envelopment analysis is in actuality there is no obvious functional form so the development of flexible level sets is beneficial. Extensive work

has been done by Simar, see for example Simar and Wilson [1998], on this data generation problem; however, fundamentally this issue is not resolved.

The comparison of various flavors of DEA, deterministic regression models, and stochastic frontier models applied to real (versus generated) data have been widely reported in the literature. Here a brief summary of some of the most interesting papers. Banker, Conrad and Strauss [1986] compared corrected least squares method with an underlying translog function to DEA's variable returns to scale (VRS) model. They find the translog model fails to reject constant returns to scale assumption while DEA identifies significant returns to scale. Also technical efficiency estimates from the translog model were less correlated with capacity utilization than the DEA efficiency estimates.

Another commonly cited paper is Hjalmarsson, Kumbhakar and Heshmati [1996] which compares DEA, SFA, and deterministic parametric frontier approach (DFA). A variety of specifications for each model are used including both variable and constant returns to scale DEA models with consideration for intertemporal and sequential models. For the SFA models three situations are considered, a standard panel data model, a model with additional control variables, and a model where the additional control variables are determinants of the technical efficiency. They found that efficiency estimates vary within models across time as much as across models. For one data set, all models showed decreasing efficiency over time and the SFA supported the assumption constant returns to scale while DEA showed evidence of variable returns to scale. Cummins and Zi [1998] compared SFA, DEA-VRS, individual specific effects (ISE), and Free-Disposal Hull (FDH) methods for the insurance industry. They concluded estimated efficiencies

averaged over time varied widely over estimation methods and results will be significantly affected by the selection of the evaluation method.

Other interesting papers that compare methods are Bauer et al. [1998], Bojanic, Caudill and Ford [1998], Coelli and Perelman [1999], and Drake and Simper [2003]. It is interesting to note the technique shown to be superior is not consistent. Because there is no clearly superior method for measuring efficiency, Charnes, Cooper and Sueyoshi [1988] are often cited with the comment, cross-checking the results of several methods is advocated for gaining greater understanding of the efficiency of observations in a sample population.

Of particular interest are the papers in which external, environmental, or uncontrollable variables are considered. This is because these models are the first steps towards explaining in more detail the sources of technical inefficiency, and hence are the starting point for developing models that will give more insight into how to improve operations. To begin, Yu [1998] compares CRS-DEA, VRS-DEA, and stochastic frontier for both the one-stage and two-stage methods, and BM environmental variable model. He used generated data and the presence of an environmental variable to see which method could most closely return the specified efficiency. He allows the magnitude of the environmental variable to vary and showed the larger the variation, the worse the models performed. His conclusion is the one-stage stochastic frontier performed the best, but noted his design of experiment may have influenced this result. Also the previously mentioned Ruggiero [1998] paper compares the BM environmental variable model, Ray's two-stage model, Lovell-Ruggiero exogenous variable model and Ruggiero's environment measurement model. He finds BM environmental variable model under-

estimates the efficiency by 19% on average, and Ray's two-stage model under estimates the average efficiency by 7%, and Lovell-Ruggiero exogenous variable model over estimates the average efficiency by 8%. However, Ruggiero's environment measurement model consistently over-estimates the average by only 3%. In the generation procedure 12.5% of DMU's were given to be efficient, and the remaining 87.5% had their inputs scaled by $1/\gamma$, where $\gamma = \exp(-|\mu|)$ where μ is normal distributed with a mean 0 and standard deviation of 0.3 .

Because technical efficiency is not directly measurable, it is hard to say, which method is superior at calculating actual efficiency. There would seem to be four possible results of comparing any two methods. First both methods could be correct, or method A could be correct and method B incorrect, or vice-versa, or both methods could be wrong. In the case of comparing DEA and SFA the result must be one of the latter three, because the efficiency estimates generated by the two methods are so different. Also even if neither method could calculate actual efficiency, there is not a clearly better method available, hence, one of these two methods should be chosen. Two papers that compare DEA and SFA are Ondrich and Ruggiero [2001] and Bauer et al. [1998]. Perhaps the most critical evaluation of SFA is a comparison using generated data performed by Ondrich and Ruggiero [2001]. They find that, because the assumed shape of the error distributions is used to identify a key production function parameter, the stochastic frontier models, like the deterministic models, cannot produce absolute measures of efficiency. Moreover, they show that rankings for firm-specific inefficiency estimates produced by corrected ordinary least squares do not change from the ranking of the

composed error SFA model. As a result, the performance of the deterministic models is qualitatively similar to that of the stochastic frontier model.

Bauer et al. [1998] introduced six consistency conditions any method of efficiency evaluation should meet. They are

i) The efficiency estimates generated by the different approaches should have comparable means, standard deviations, and other distributional properties (skewness would be another good candidate);
ii) The different approaches should rank the DMUs in approximately the same order;
iii) The different approaches should identify mostly the same institutions as "best-practice" and as "worst-practice";
iv) All of the useful approaches should demonstrate reasonable stability over time, i.e., tend to consistently identify the same institutions as relatively efficient or inefficient in different years, rather than varying markedly from one year to the next;
v) The efficiency estimates generated by the different approaches should be reasonably consistent with competitive conditions in the market; and
vi) The measured efficiencies from all of the useful approaches should be reasonably consistent with standard non-frontier performance measures, such as return on assets or the cost/revenue ratio.

Bauer et al. [1998] compare four different methods: DEA, SFA, ISE, and thick frontier approach (TFA). Their results show the parametric methods (SFA, ISE and TFA) satisfy the first three conditions among themselves. However, DEA does not give results consistent with the other methods. On the other 3 conditions they find the parametric methods to satisfy the conditions, but they find DEA fails condition v and vi.

The six consistency conditions proposed seem to be a start at developing an axiomatic approach to determine which efficiency evaluation method produces more appropriate results. However, several of the conditions seem misstated. Condition iv states persistence should be a valuable characteristic of an efficiency measure. Similar to forecasting, techniques to measure efficiency are often based on slow moving (slow to change) variables. When this is true, efficiency estimates should be persistent. However,

efficiency estimate persistence is not a valuable characteristic. It would be preferred for an efficiency measure to change when pertinent variables change, for example, management or ownership changes or market conditions change. Condition v states "efficiency scores generated by the different approaches should be reasonably consistent with competitive conditions in the market", however, a method for quantifying 'competitive conditions in the market' is not stated. Rather this condition is used in an ad-hoc manner to allow for subjective expert opinion. The idea behind condition v seems to be desirable; however, a rigorous measurement needs to be defined in order to quantify, measure and compare the concept. While vi (measured efficiencies from all of the useful approaches should be reasonably consistent with standard non-frontier performance measures) is a condition that would be desirable, a warning is needed. Many of the non-frontier performance measures relate to profit, which is significantly influenced by market conditions. One point related to this issue is made very nicely by Mr. Sturrock in his comments on Farrell [1957],

> "…the average cost curve, as output increases in a business, is normally U-shaped. Average cost per unit first falls owing to economies of scale, then rises as one reaches the full capacity of the existing equipment. The optimum level of production, however, is not the point of lower average cost. In practice, it pays to increase production until marginal cost has risen to marginal revenue."

This concept should serve as a warning or reminder that additional assumptions about market conditions and firm behavior are needed for non-frontier performance measures to be highly correlated with efficiency.

While the most common methods to handle exogenous variables is to use the BM environmental variable model or Ray's two-stage model, other more complicated models have been proposed in the literature. Two such models are the four-stage method and the

three-stage method. A four stage method was demonstrated in Fried, Schmidt and Yaisawarng [1999] to adjust DEA estimates for environmental factors. The authors use both DEA and Tobit regression to separate the components of efficiency related to environmental factors and managerial ability. The 1999 paper also introduces an interesting technique of summing the radial and non-radial input slack for an input-oriented model or similar values for an output oriented model. This allows not only DMUs falling on the interior of the frontier to be inefficient, but also DMUs on the free disposability portions of the frontier. Fried et al. [2002] shows how to combine DEA with SFA in a three-stage process to separate managerial efficiency, environmental effects and random noise. The procedure is very similar to the 1999 model, but by using SFA with the two-part error component, the 2002 model allows random noise to be separated from managerial efficiency.

Stochastic production frontier approach (SFA) and data envelopment analysis (DEA) have been used to quantify technical inefficiency under a variety of assumptions or situations, Fried, Lovell and Schmidt [1993], but there is still no clear connection between these evaluation methods and specific advice to individual production units about improvement methods or strategies.

3.6 Real vs. generated data

Not much is known about the actual unobservable efficiencies measured from the feasible efficient frontier. This is why it is possible for DEA and SFA to generate very different results and no one can state definitely which is correct. This also calls into question the validity of any data generation process. Although many people attempt to

generate data there is no way to know if the data could represent an actual group of systems.

Because DEA is a frontier method, it is very sensitive to extreme values for inputs and outputs. Using DEA with standard data generation techniques, which assume a distribution and generate data based on that distribution, tends not to work well. This is because the results are very sensitive to the random occurrences of a draws from the tails of these distributions. Data generation techniques for DEA are closely tied to progress in the field of rare event simulation, which has been slow. Thus, it is strongly recommended that DEA and other frontier methods be performed on real data sets. While this may limit results to being specific to a particular industry or system in the short-run, this is the only way to develop the data and the insight to allow a better understanding of the inefficiency in long-run.

3.7 Peer group research

A critical component of any analysis is identifying the members of the group used as the reference for analysis. This is often taken to mean selecting an industry and deciding which companies are in the industry. However, it should be noted in productivity analysis, particularly when non-parametric methods are used, it is very important to identify a group of peers that are not only using similar inputs and outputs, but are using comparable production technologies. One implied assumption is that any unit under consideration could switch its current technique for production to the technique of any other unit in the peer group. If this is not true the analyst should question if the units are truly from the same peer group.

An example of this problem is given by Thore, Kozmetsky and Phillips [1994], who analyzed the computer industries and included computer companies such as Apple Computer, Atari, Compaq Computers, and Cray Research. While all being identified as being in the same industry these companies use very different production technologies. For example Atari specialized in home gaming systems and software development while Cray Research specializes in large mainframes sold primarily to universities or large corporations. It is hard to imagine very much in common between these two production processes, or customer groups.

Although there appears to be no literature on the general question of how to identify a group, there are several papers which describe their specific methods for selecting a group consistent with the criteria stated above. In McCarty and Yaisawarng [1993] the problem is to decide if the plaintiff schools in the case Abbott v. Burke are poorly managed as claimed by the state of New Jersey. The approach is to limit the peer group to the 28 court-identified districts so that the results will support policy decisions involving those districts. By creating this subgroup the authors imply that the production technology of education in "poorer, urban" school districts may be different than the production technology used in other districts in New Jersey.

Similarly, in Fuentes, Grifell-Tatje and Perelman [2001], the definition of the group to be analyzed was carefully considered. Their paper is concerned with measuring the rates of growth and technical change in the Spanish insurance industry. There are three branches of the insurance market: health, life and non-life. They chose to analyze only companies operating in all three branches. In a footnote they mention it could be argued that the decision to operate in all three branches is not based on efficiency considerations,

in which case a selection bias is avoided. However, in the case of the Spanish insurance companies their decision to be specialized in one branch or in a combination of particular branches reflects historical and market considerations.

One final example of group identification is given in Charnes, Cooper and Rhodes [1981]. Here they are analyzing the effects of Program Follow Through in which continued help is provided to disadvantaged students who were in the Project Head Start. Project Head Start was a project for preschool students, and Program Follow Through was suggested as a continuation to give the same students assistance from kindergarten until the third grade. Only certain students from Project Head Start were allowed to participate in Program Follow Through. Here again there are natural groups for analysis, Program Follow Through members and students that participated in Project Head Start, but not Program Follow Through. In this case the authors were able to gather data on these two groups and compare their performance to decide if Program Follow Through was worth the investment. Their comparison of the two groups used visual inspection to draw the conclusion, for certain mixes of inputs Program Follow Through was successful, but in other regions or mixes of inputs Program Follow Through performed worse than the no program case. The conclusion was the program was not worth the investment. It is often difficult to define groups; however, from this small group of examples the results generated are based on more realistic comparisons.

3.8 Data requirements research

In describing a data set to be used in a DEA model there are typically three variables of concern: number of inputs (p), number of outputs (q), and number of data points or DMUs (n). It is valuable to understand the relationship between these three parameters.

Generally, one would expect that larger model (larger sum of $p+q$) specifications would need larger data sets (larger n) to achieve valid results since each additional input and output would increase the dimensionality and complexity of the production frontier. However, the concept of a strict lower bound on n does not seem like a useful concept. Rather a person interested in using DEA, gains increased assurance the measure of efficiency given from the model is approaching the actual efficiency as the number of DMUs used in the reference set increases. Hence the question of greater interest is "how many DMUs are required for the efficiency estimates generated to be reasonable approximations of actual efficiency?" The first reference to this problem is made in Banker et al. [1989]. Here the authors state,

> "Efficiency evaluations associated with these solutions will be dependent on the number of degrees of freedom that are available. There are p + q constraints to be satisfied in the problem (CRS DEA) and n observations, one for each of the j = 1,…,n DMUs that form the possible combinations from which efficiency evaluations can be secured. From degrees of freedom considerations, the number of variables λ_j used for the solution in the CRS DEA should be at least as great as the number of constraints. Thus, the number of DMUs for which there are observations should be greater than the number of constraints and, for DEA efficiency evaluations; it is generally advisable to have
>
> $$n \geq 3(p+q)$$
>
> This is only a rule of thumb, of course, which may need to be adjusted in particular situations."

This argument is based on the number of non-zero λ_j; however, it is possible to argue base on discriminating power for a different value. This is what Boussofiane, Dyson and Thanassoulis [1991] did stating,

> "The selection of inputs and outputs can affect the discriminating powers of DEA as the number selected needs to be small compared to the total number of units for effective discrimination. The total number of such ratios (output to input) will be the product of the number of inputs and outputs and this product is a reasonable indicator of the minimum number

> of efficient units…so that the total number of units in the set needs to be much greater…".

This implies

$$n >> (p * q)$$

Since $p*q$ grows much more quickly than $p+q$, the two recommendations can differ dramatically. Cooper, Seiford and Tone [2000] merge the two previous criteria giving the following reasoning,

> "As in statistics or other empirically oriented methodologies, there is a problem involving degrees of freedom, which is compounded in DEA because of its orientation to relative efficiency. In the envelopment model, the number of degrees of freedom will increase with the number of DMUs and decrease with the number of inputs and outputs. A rough rule of thumb, which can provide guidance, is as follows.
>
> $$n \geq \max\{p \times q, 3(p+q)\}$$
>
> where n = number of DMUs, p = number of inputs and q = number of outputs."

Dyson et al. [2001] attempt to make Boussofiane et al.'s recommendation slightly more rigorous with the following statement,

> "A suggested 'rule of thumb' is that, to achieve a reasonable level of discrimination, the practitioner needs the number of units to be at least 2pxq where pxq is the product of the number of inputs and number of outputs."

Anderson and Hollingsworth [1996] also explore the question of how many data points are needed for varying size models. They designed an experiment where data was generated with a known average efficiency level. Recognizing that DEA estimates are upwardly biased in small samples, data points were added to the analysis until the average efficiency over all units in the analysis was within a stated percentile of the generated average efficiency level. By the average sample efficiency and the generated

efficiency levels being close, the proximity of the two frontiers was assumed to be close. They find 140 data points are needed for the average DEA estimate to be within the 95th percentile of the generated average efficiency level for complex models $(p+q>9)$.

The question of how much data is required still has not been answered. Anderson and Hollingsworth made some progress, but the criterion, frontier proximity, is measured simply by averaging the DEA estimates and comparing the actual mean of the generated data. The average of the DEA estimates is a number, but it is not related to the complexity or shape of the production frontier. If the goal of sampling DMUs for analysis is to take a sample that will generate a production frontier similar to the entire set, then a more specific method for measuring the proximity of two frontiers is needed. This topic will be further explored in Chapters 4 and Chapter 7.

3.9 Warehouses

3.9.1 The warehouse's role in a logistic system

Industrial engineers study a variety of integrated systems and their components to include, but not limited to, manufacturing, logistic, agricultural, military, health, and service systems. The implementation of the methods developed here will focus on distribution centers, which are components of logistic systems, but are complicated systems in their own right, bringing together people, material, information, equipment, and energy. Of the systems listed above, military and health systems often contain logistic systems while manufacturing, agricultural and service systems create products that flow through logistic systems. Warehouse systems were selected for this study because in some sense they are the most basic of the integrated systems. By being part of

the whole logistic system, a warehouse must be less complicated than a logistic system. Also a warehouse does not create any new products, rather warehouse receive, store, gather, pack, and ship items. Other systems need to perform all these same tasks and create products; therefore warehouse are simpler systems.

Define a *supply chain* as the linked set of resources and processes that begin with the sourcing of raw material and extends through the delivery of end items to the final customer. Then a *logistic system* is the transportation and material-handling network that underlies a supply chain. Components of both the logistic system and supply-chain are vendors, manufacturing facilities, transportation companies, internal warehouse, distributors, wholesalers and all other entities that lead up to final customer acceptance. A *warehouse or distribution center*, is a location used in the process of marketing and supplying goods to members of the supply chain, typically occupying a large building and using specialized equipment to store goods temporarily.

The purpose of the distribution center is three fold: to realize economies of scale in production and shipping, to consolidate product, and to reduce response time to customers. While implementing particular warehousing methods, one purpose may be sacrificed to improve another. Economies of scale implies that a company owning a distribution center can order in larger quantities from their suppliers or make larger requests to their upstream within-firm manufacturers. In exchange for the larger order, which guarantees demand for the supplier and allows them to engage in longer production runs, the supplier is often able to give the customer a quantity discount or a price break (charging them less per item than if the customer had placed a smaller order).

Similarly, the within-firm manufacturer can amortize large setup costs for production over more units, reducing the cost associated with each unit.

Retail outlets often are located in areas where people like to live and where other businesses are located for cross selling effects. This causes land to be expensive and competition for labor to increase wages. Therefore, stocking and handling inventories at a retail outlet is costly. A distribution center allows a firm to consolidate inventories of several retail outlets in one location, therefore reducing the effects of random fluctuations in demand at a single outlet. The distribution center does not need to be in a particularly populous area and hence land and labor costs are lower. Finally by shipping to a distribution center before the items are taken to a retail outlet, the firm can amortize the fixed cost of transportation over large shipment sizes and over the longer distances. For example having a distribution center may allow the firm to order in truckload quantities, which is significantly cheaper than less-than-a-truckload on a per volume basis. The truckload can then be taken from the manufacture to the distribution center and from the distribution center relatively short deliveries can be made to retail outlets. Often this method is significantly cheaper than small shipments sent from the manufacture to each retail outlet.

Distribution centers allow small quantities of inventory to be kept at local retail outlets even though manufacturing is in relatively far away places. When a retail outlet needs more of an item, having a distribution center often helps to reduce response time. When the retailer realizes the outlet's inventory is low they often want to place a replenishment order. If the manufacturer is far away or does not keep a finish units inventory, this may mean the retailer will not be able to get a replenishment order before

the remaining outlet inventory is sold or the selling opportunity has passed. A distribution center allows for replenishment orders without an additional production run and often reduces the lead-time to deliver replenishment orders.

3.9.2 Warehouse description

A distribution center is sometimes referred to as a cost center because the activities within the facility do not generate revenue, but are a necessary step in getting the goods to the customer. The costs come from inventory holding cost because there is money tied up in physical items sitting on the shelves and significant amounts of money are spent to store and access inventory. The distribution center is often asked to provide services such as storing materials so economies of scale can be realized elsewhere, consolidating products so customers (or retail outlets) can receive the mix of products they would like in fewer shipments, and providing quick response times while minimizing the costs associated with carrying inventory and handling inventory. To this end there are four major functions within a warehouse: receiving, storage, order assembly and shipping.

Each function is made up of a series of activities. Within receiving the activities of unloading, inspecting and putting away take place. The only activity in the storage function is the storage activity. The assembly function consists of picking and grouping items by the orders in which they are requested. Finally, the shipping activity includes packing and loading.

There is a wide variety of equipment that can be used to store items. In general there are various types of racking, there are automated storage and retrieval systems (ASRS), and items can simply be stacked on the floor. Typically, distribution centers do not change their storage equipment often. Thus the other equipment in the facility is often

chosen to be compatible with the storage equipment. Typical equipment is pallet jacks, lifts trucks, conveyors and others.

The activities within the warehouse are in some cases done simultaneously; while in other cases may be divided by shift. For example some distribution centers do all picking during first shift and all receiving during second shift or receiving in the morning and shipping in the afternoon. This can help reduce congestion and simplify management. The characteristics of a warehouse will be further discussed in Chapter 7.

To apply efficiency measurement methods to warehousing, an input / output model needs to be specified. The inputs selected are labor, investment, and space. Labor is measured as annual labor hours including both direct and indirect labor to perform necessary operations of receiving, moving, storing, retrieving, order picking and shipping. Some indirect labor, such as management, planning, and equipment maintenance, are included. However, indirect supporting personnel, such as security, cleaning staff, office assistants, accounting, human resources, customer service, and the labor assigned to the value-adding activities, are not counted.

Investment is measured by taking an inventory of the equipment used in the warehouse and assigning standard values measured in U.S. dollars, regardless of its age, then multiplying the standard values versus the quantity and summing over all equipment types. Space is the area measured in square feet, dedicated to the warehouse operations of receiving, put away, storing, retrieving, order picking, packing and shipping. Areas for supporting activities, such as offices, rest rooms, cafeteria, or break rooms are not included.

To measure the productivity of the distribution center the inputs consumed and the outputs generated from the activities stated above need to be identified and quantified. Resources (inputs) that are consumed by activities that may be collocated with the distribution center, but do not fall within the activities stated will be excluded from the analysis. The initial model used is a 5 output by 3 input model. Warehouses actually use other inputs and outputs, but this set of 8 measures is believed to capture the most important inputs used and outputs generated. This model was initially developed by Hackman et al. [2001] and used on a data set of approximately 50 warehouses.

The outputs are broken case lines shipped, full case lines shipped, pallet lines shipped, accumulation, and storage function. When items arrive at a warehouse, they typically come in cases stacked on a pallet. Depending on the customer types and demand patterns warehouses will receive orders requesting a certain number of each holding size: pallets, cases, or items contained in the case.

An order from a customer is made up of lines. Each line is particular to a sku number, or in other words, it is particular to a certain item in a certain holding size. Thus broken case lines shipped are the number of broken case lines summed over all shipped orders for the 12-month period in which data was collected. Similar definitions are true for full case and pallet lines shipped. Accumulation is the difference between lines shipped summed over all holding sizes and the number of orders shipped. Thus accumulation is not a ratio, but is a scalar which characterizes the effort that the warehouse makes to consolidate lines picked for the same customer order.

Finally, storage function is a number intended to describe the storage effort of the warehouse. Part of the purpose of the warehouse is to hold inventory so that upstream

producers can have longer production runs, while downstream customers can receive quicker responses to their requests for goods. The formula for storage function is found in Hackman et al. [2001]. These five outputs are the most important outputs for our set of warehouses.

Distribution centers are often members of larger supply chains or networks. As a member of the network, the number of shipments demanded from a distribution center and the number of replenishments received at a distribution center are often affected or even controlled by supply chain coordination efforts. Therefore, when output is measured in lines shipped or service provided, those numbers often are controlled outside of the distribution center. This suggests that analysis of the productivity should take an input orientation because relative to outputs, the distribution center has more control over the level of inputs.

3.10 Conclusion

A variety of parametric and non-parametric models have been developed to quantify efficiency and measure performance. The non-parametric method distinguishes itself by its ability to model production behavior without assuming a functional form of the DMU objectives, the production relationships, or the statistical distribution of the inefficiencies. However, each non-parametric model imposes a set of assumptions with regard to the production possibility set, returns-to-scale properties and the similarity of the peer group. The selection of the appropriate model involves a difficult trade-off between small sample error and specification error. Using more general production assumptions can substantially reduce discriminating power, especially in small samples. On this basis model specification and data consistency tests are valuable contributions to the literature

because the results of these tests allow analysts to better understand the quality of the assumptions imposed. While there are a wide variety of models proposed in the productivity literature, the lack of statistical specification tests for the underlying assumptions of the standard non-parametric models is an opportunity for further research.

Understanding the impact of environment and practices of the DMU under consideration has often been modeled as an after thought of efficiency measurement. Ruggiero [1996] attempts to address the impact of environment and practices while measuring efficiency. Ruggiero makes a similar observation as Hall and Winsten [1959], noting that DMU operating in different environments may not be compared fairly. The implication of this fact makes the current methods of non-parametric performance measurement invalid. The basic assumption of non-parametric performance measurement is all DMUs are using the same technology, and the behavior of one DMU can be mimicked with similar results by a different DMU. This assumption is contradicted by Hall and Winsten's observation. This realization provides strong motivation for developing techniques to identify subsets of the data for which the basic non-parametric assumption holds, e.g. the practices of one DMU can be adopted by another member of the subset with similar results.

Another model is available for quantifying the impact of environment and practices. The two-stage model is a reasonable approach to the extent that all observations are comparable, and practices and attributes represent a common set of choices available to all DMUs. The second stage regression identifies correlations between decisions made and observed efficiency given the choices of one DMU can be repeated by a different DMU with similar input usage and output production. Whereas Hall and Winsten's

observation is not consistent with the assumption of a common technology used by all DMUs, the two-stage method is consistent. After the analyst accepts the assumption of a common technology, to use the two-stage method, the analyst needs to further assume

1) the practices and attributes influence the input and output levels, but are not substitutes for inputs or outputs
2) and DMUs can adjust their practices and attributes

When the relevant attributes are related to the environment or long-term decisions, the second assumption of the two-stage model becomes more tenuous.

The two-stage model and the Lovell-Ruggiero model are the two most common models used to investigate practices and attributes. However, the assumptions related to both models have been described and contrasted. Hall and Winsten's concern could be addressed by a model that identified comparable peer groups perhaps based on environmental variables or other data, and then applied the two-stage method. This model could maintain a large enough comparison group to use non-parametric methods with their limited imposed restrictions.

CHAPTER 4

AN OUTLIER DETECTION METHODOLOGY WITH CONSIDERATION FOR AN INEFFICIENT FRONTIER

4.1 Introduction

Productivity and efficiency have been research areas for both economists and engineers for the past fifty years. Productivity is the ratio of outputs produced to inputs consumed and efficiency is the ratio of a given system's productivity compared to the best possible productivity, Lovell [1993]. Many models have been proposed for determining the best possible productivity. A main concern while constructing these models or evaluating them is deciding if the productivity identified is truly achievable for the system under consideration. This has lead researchers to investigate and quantify the effects of the environment and other variables that cannot be controlled by system management. One of the most common types of models for this purpose has come to be known as the two-stage semi-parametric models, first suggested by Timmer [1971].

In the first stage a deterministic frontier model is constructed. When the assumptions of convexity and free disposability are made, this calculation is referred to as data envelopment analysis (DEA), made popular by Charnes, Cooper and Rhodes [1978]. Other deterministic frontier techniques also may be used, such as free disposal hull (FDH), first rigorously analyzed by Deprins, Simar and Tulkens [1984].

In the second stage the efficiency estimates calculated in the first stage are regressed against a variety of environmental variables. The first implementations of the two-stage semi-parametric models were by Ray [1988] and Ray [1991]. However, these methods

have recently been criticized by Simar and Wilson [2005] for their lack of a coherent data-generating process and mishandling of the complicated unknown serial correlation among the estimated efficiencies.

Wilson [1995] and others note that in the first stage the deterministic nature of the frontier means errors in measurement in the observations supporting the frontier could cause severe distortions in the measures of efficiency for the entire population. Wilson then suggests a method to remedy this problem by calculating the leave-one-out efficiency, sometimes called super efficiency or jackknife efficiency, and identifying outliers based on the leave-one-out efficiency estimate. The leave-one-out efficiency estimate has been presented by Banker, Das and Datar [1989], Anderson and Petersen [1993], and Lovell, Walters and Wood [1993] among others. The Banker paper refers to its use for outlier measurement whereas the latter two papers use the method for tie breaking among the observations that appear to be efficient. Wilson then relates this problem of identifying observations with measurement error to the problem of outlier detection in the classical linear regression models. However, outliers in linear regression models can be found both above and below the regression line, whereas, Wilson's method only identifies a subset of outliers related to being "too good" or to continue the regression analogy, outliers found above the regression line.

While outliers are an intuitive concept, a rigorous definition is hard to state. Assuming data have been generated by drawing from a distribution, an observation categorized as an outlier may represent a low probability draw (for example a draw from one of the tails in the normal distribution). While this may appear to be an outlier, as Cook and Weisberg [1982] point out, this type of observation may lead to the recognition of important phenomena that might otherwise go unnoticed. With this in mind, the rather loose definition of outlier provided by Gunst and Mason [1980], "as observations that do not fit in with the pattern of the remaining data points and are not at all typical of the rest

of the data", seems appropriate. In deterministic frontier models, outliers that support the frontier can be thought of as observations that are "too good" and thus are particularly dangerous, as noted above by Wilson. The observation motivating this work is: *when the two-stage semi-parametric model is used, outliers that represent particularly bad performance might distort the second stage results*.

There has not been much research in the area of identifying outliers relative to a nonparametric deterministic frontier. There appears to be no published literature discussing how to identify outliers which distinguish themselves by having particularly poor performance. The available research (Wilson [1995] and Simar [2003]) focuses only on identifying outliers which impact the efficient frontier. Many studies have been performed to measure sensitivity or robustness of DEA results and while this is closely related to many techniques for identifying outliers, the concept is fundamentally different.

There has been limited attention paid to inefficient frontiers. Paradi, Asmild and Simak [2004] suggest a worst practice detection method by applying traditional DEA models when only detrimental (bad) outputs are selected. In their approach, a new mathematical formulation is not needed; poor performers are simply identified by high levels of bad outputs. Liu and Hsu [2004] also have suggested similar mathematical formulation for identifying an inefficient frontier; however, the paper provides no motivation for developing an inefficient frontier.

The present chapter describes an inefficient frontier and how this concept can be used to identify outliers that distinguish themselves by having particularly poor performance. These observations should then be examined further to determine if an error has taken place, possibly in data entry or in identifying these units as members of the peer group for

this analysis. This chapter also describes the implementation of the iterative outlier identification process, discussed in Wilson [1995] although apparently not demonstrated in the literature.

Section 4.2 will review the two-stage semi-parametric models using data envelopment analysis (DEA) in the first stage and bootstrapping methods in the second stage. Section 4.3 will address methods for constructing an inefficient frontier and describe outlier detection methods applied to the inefficient frontier. An example using the classic Banker and Morey [1986b] data set will be shown in section 4.4. The impact on second stage results of not identifying and processing inefficient outliers will be demonstrated. Finally, conclusions will be presented.

4.2 Description of Two-Stage Semi-Parametric Bootstrapping Method

The two-stage semi-parametric model approach consists of estimating efficiencies in the first-stage and regressing these efficiency estimates against a set of environmental variables in the second-stage. Many models are available for estimating efficiency; we will focus on the DEA model. The DEA production set can be described by

$$\hat{P} = \left\{ (x, y) \mid y \le Y\lambda, x \ge X\lambda, i^T\lambda = 1, \lambda \in R_+^n \right\} \tag{4.1}$$

where $\hat{P}$ is an estimate based on the observed pairs (x_i, y_i) of the actual production set P, $x \in R_+^p$ denotes a $(1 \times p)$ vector of inputs, $y \in R_+^q$ denotes a $(1 \times q)$ vector of outputs, n is the number of observations, $Y = [y_1 ... y_n]$, $X = [x_1 ... x_n]$, i denotes an $(n \times 1)$ vector of ones, and λ is an $(n \times 1)$ vector of intensity variables. The production set can be completely described by either the input requirements set or the output requirement set. The input set can be stated as

$$L = \left\{ x \in R_+^p \mid \text{x can produce y} \right\} \tag{4.2}$$

and the output set as

$$K = \left\{ y \in R_+^q \mid \text{y can be produced by x} \right\} \tag{4.3}$$

To simplify exposition, we will focus on the input space; however, the concepts described for the input space transfer easily to the output space. For further description of the relationship between the two spaces see either Lovell [1994] or Charnes et al. [1993]. The linear program for calculating the efficiency estimates in the input requirement space is

$$\begin{aligned} &\min_{\hat{\theta}_{iI},\lambda} (\hat{\theta}_{iI}), \\ &s.t. \;\; -y_i + Y\lambda \geq 0, \\ &\quad \hat{\theta}_{iI} x_i - X\lambda \geq 0, \\ &\quad \sum_{j=1}^{N} \lambda_j = 1 \\ &\quad \lambda_j \geq 0 \end{aligned} \tag{4.4}$$

This linear program is solved once for each observation, $i = 1...n$ to compute efficiency estimates for the observation.

Let $z \in R_+^r$ denotes a $(1 \times r)$ vector of environmental variables. In two stage analysis, a function, typically $\psi(z_i, \beta) = z_i\beta$ is specified and an associated regression model is

$$\hat{\theta}_i = z_i\beta + \varepsilon_i \tag{4.5}$$

where $\hat{\theta}_i$ are the efficiency estimates from the first-stage with the subscribe I dropped with the understanding either the input oriented efficiency estimates can be used or $1/\hat{\theta}_{iO}$ the output oriented efficiency estimates. The sign on the resulting coefficients indicate the direction of the influence and hypothesis testing can be performed to assess the

significance. Until recently this had been standard practice and advocated by several researchers, e.g., Coelli, Rao and Battese [1998], McCarty and Yaisawarng [1993], and Ray [1991].

In 2005 Simar and Wilson introduced a bootstrapping technique to replace the second-stage regression. They cited the need for a new technique as two fold: 1) the original two-stage method lacks a coherent data-generating process; and 2) it mishandles the complicated unknown serial correlation among the estimated efficiencies and the correlation between the ε_i and the z_i.

The Simar and Wilson bootstrapping technique uses Shephard's input distance function which is inversely related to an input efficiency estimate. Shephard's input distance function is

$$D_{iI}(x_i, y_i) = \left(\hat{\theta}_{iI}\right)^{-1} = \max\left\{\hat{\delta}_{iI} \middle| \frac{x_i}{\hat{\delta}_{iI}} \in L(y)\right\} \tag{4.6}$$

The value of $\hat{\delta}_{iI}$ is a normalized measure of the distance from a point (x_i, y_i) to the frontier, holding output levels and the direction of the input vector fixed. For completeness and later use we also introduce Shephard's output distance function here as

$$D_{iO}(x_i, y_i) = \min\left\{\hat{\delta}_{iO} \middle| \frac{y_i}{\hat{\delta}_{iO}} \in K(x)\right\} \tag{4.7}$$

The output distance function D_{iO} gives a normalized measure of the distance from a point (x_i, y_i) to the frontier, holding input levels and the direction of the output vector fixed.

Next Simar and Wilson suggest the following algorithm for the second-stage (in the following algorithm the I on $\hat{\delta}_{iI}$ is dropped to simplify the notation):

[1] Using the original data, compute $\hat{\delta}_i = \hat{\delta}\left(x_i y_i \middle| \hat{P}\right) \forall i = 1,...,n$ from (4.4) and the inverse relationship between efficiency estimates and Shephard's distance function.

[2] Use the method of maximum likelihood to obtain an estimate $\hat{\beta}$ of β as well as an estimate of $\hat{\sigma}_\varepsilon$ of σ_ε in the truncated regression of $\hat{\delta}_i$ on z_i in $\hat{\delta}_i = z_i\beta + \varepsilon_i \geq 1$.

[3] Loop over the next four steps ([3.1]-[3.4]) L_1 times to obtain n sets of bootstrap estimates $B_i = \left\{\hat{\delta}^*_{ib}\right\}_{b=1}^{L_1}$:

[3.1] For each i = 1,…,n draw ε_i from the $N(0, \hat{\sigma}_\varepsilon^2)$ distribution with left truncation at $\left(1 - z_i\hat{\beta}\right)$.

[3.2] Again for each i = 1,…,n compute $\delta_i^* = z_i\hat{\beta} + \varepsilon_i$.

[3.3] Set $x_i^* = x_i, y_i^* = y_i \frac{\hat{\delta}_i}{\delta_i^*}$ for all i = 1,…,n.

[3.4] Compute $\hat{\delta}_i^* = \hat{\delta}\left(x_i y_i \middle| \hat{P}^*\right) \forall i = 1,...,n$ where $\hat{P}^*$ is obtained by replacing Y, X X in (4.4) and the inverse relationship between efficiency estimates and Shephard's distance function.

[4] For each i = 1,…,n compute the bias-corrected estimator $\hat{\hat{\delta}}_i$ defined by $\hat{\hat{\delta}}_i = \hat{\delta}_i - \text{BIAS}\left(\hat{\delta}_i\right)$ where

$\text{BIAS}\left[\hat{\delta}(x,y)\right] = B^{-1}\sum_{b=1}^{B}\hat{\delta}_b^*(x,y) - \hat{\delta}(x,y)$ were $\hat{\delta}_b^*(x,y)$ are the bootstrap estimates from step [3.4] and $\hat{\hat{\delta}}_i$ is the original estimate.

[5] Use the method of maximum likelihood to estimate the truncated regression of $\hat{\hat{\delta}}_i$ on z_i, yielding estimates $\left(\hat{\hat{\beta}}, \hat{\hat{\sigma}}\right)$.

[6] Loop over the next three steps ([6.1]-[6.3]) L_2 times to obtain a set of bootstrap estimates $C = \left\{\left(\hat{\beta}^*, \hat{\sigma}_\varepsilon^*\right)_b\right\}_{b=1}^{L_2}$:

[6.1] For each i = 1,…,n draw ε_i from the $N\left(0, \hat{\hat{\sigma}}\right)$ distribution with left-truncation at $\left(1 - z_i\hat{\hat{\beta}}\right)$.

[6.2] Again for each i = 1,…,n compute $\delta_i^{**} = z_i\hat{\hat{\beta}} + \varepsilon_i$.

[6.3] Use the maximum likelihood method to estimate the truncated regression of δ_i^{**} on z_i, yielding estimates $\left(\hat{\hat{\beta}}^*, \hat{\hat{\sigma}}^*\right)$.

[7] Use the bootstrap values in C and the original estimates $\hat{\hat{\beta}}, \hat{\hat{\sigma}}$ to construct estimated confidence intervals for each element of β and for σ_ε.

The confidence intervals can be constructed for any of the parameters including the efficiency estimate in the following manner. To find a confidence interval for a particular β_j consider the value $\left(\hat{\hat{\beta}}_j - \beta_j\right)$. A confidence interval for β_j can be defined as

$$\Pr\left[-b_{\alpha/2} \le \left(\hat{\hat{\beta}}_j - \beta_j\right) \le -a_{\alpha/2}\right] = 1-\alpha \tag{4.8}$$

If the distribution of β_j were known then it would be straightforward to find a_α, b_α. However, it is not, so the following bootstrap assumption is used:

$$\left(\hat{\hat{\beta}}_j - \beta_j\right) \Big| P \overset{approx}{\sim} \left(\hat{\hat{\beta}}_j^* - \hat{\hat{\beta}}_j\right) \Big| \hat{P} \tag{4.9}$$

Thus equation (2.8) is approximated by

$$\Pr\left[-b^*_{\alpha/2} \le \left(\hat{\hat{\beta}}_j^* - \hat{\hat{\beta}}_j\right) \le -a^*_{\alpha/2}\right] \approx 1-\alpha \tag{4.10}$$

Based on the bootstrapping results, an empirical distribution for $\left(\hat{\hat{\beta}}_j^* - \hat{\hat{\beta}}_j\right)$ can be constructed and equation 2.10 can be applied to the empirical distribution in order to find $a^*_{\alpha/2}$ and $b^*_{\alpha/2}$. The values of L_1 and L_2 need to be specified in order to implement the bootstrap in the second-stage. Simar and Wilson [2005] suggest the values of $L_1 = 100$ and $L_2 = 2000$ after testing various values. The bootstrap approximation converges to $\hat{\hat{\beta}}_j$ as the number of iterations in the bootstrap, L_2, increases.

The method suggested by Simar and Wilson improves the original two-stage model that only allowed point estimations of efficiency. By constructing confidence intervals one can begin to quantify the uncertainty related to efficiency estimates based on an

estimate of an unobservable frontier. A confidence interval provides the analyst significantly more information than the point estimate. However, before the two-stage model can be implemented, unexplainable outliers should be identified and removed from the data.

The previous outlier detection methods were only concerned with the overly efficient outliers because they were developed with the traditional deterministic frontier model in mind and did not consider the second stage regression. Thus the outliers that were overly inefficient would have minimal impact on their results. However, when the two-stage model is considered the effect of overly inefficient outliers may cause misleading results in the second stage, as will be shown in section 4.4.

4.3 The Inefficient Frontier, Outliers, and a Detection Methodology

The outlier detection methodology for non-parametric efficiency evaluation described here is distinguished from previous methodologies by incorporating a search for inefficient outliers. In order to identify inefficient outliers a standard for the lowest rational inefficiency level needs to be defined. This is done through the concept of the inefficient frontier, introduced below. The method that will be used to identify outliers is the leave-one-out method described by Wilson [1995]. The basic assumption of deterministic frontier models is all the observed units belong to the production possibility set. However, in the leave-one-out method, this assumption is relaxed to quantify the degree to which each efficient unit might be considered as an outlier. The method will be applied relative to both the efficient and inefficient frontiers and all outliers identified will be removed from the reference set, unless there is a clear argument to keep a

particular data point. After the outliers are removed any other data analysis can take place such as the two-stage semi-parametric method.

Barnett and Lewis [1995] outline four possible sources for outliers. The first, which they call deterministic, is a result of errors in measurement, recording, or understand of the value requested. Each of these leads to erroneous data values for the given observation thus the pattern generated by the other observations is not consistent with the given observation. A second source, incorrect expectations, is the result of underestimating or failing to assume the actual data pattern. For example, data may be assumed to follow a certain distribution with a given mean and variance. This distributional information is often used as the basis for developing statistical outlier tests. If the expectation of the variance is different then the true variance, an observation without deterministic outlier sources, could be identified as outliers when the observation was correctly measured.

A third source is slippage, a process that happens over time. A set of observations is taken at time t and the same operations are observed later at time $t+1$. The conditions under which the operations perform may have undergone a systematic change causing the distribution of these operations to shift. The term slippage comes from the most widely studied case in which the mean of the distribution decreases or "slips" downward over time. When all observations are examined in a single cross-section, observations from later periods are more likely to have lower means and be identified as outliers to the group as a whole. Thus a "slipping" of the mean cause an observation to be identified as an outlier.

A fourth source of outliers is contamination. Contaminates are observations belonging to a separate group from the one under evaluation. A statistician may say contaminates are observations taken from a separate data generation process. When the number of contaminated observations relative to the number of total observations is small, outlier detection techniques can be used to attempt to identify the observations belonging to a separate group. However, as the ratio of contaminated observations to total observations increases the type of problem changes from an outlier detection problem to a clustering problem. The clustering problem attempts to identify the groups present in a set of observations and assign observations to groups. Thus contamination is a reason to apply both cluster algorithms and outlier detection algorithms; however, the two problems are fundamentally different. Simar [2003] suggests performing clustering analysis before using an outlier detection method.

4.3.1 The Inefficient Frontier

Just as an efficient frontier can be calculated from observations taken from the production set P, an inefficient frontier also can be calculated. It is argued that the efficient frontier represents the maximum output given an input level, and without improvements in technology, it is not possible to achieve greater production levels. The deterministic inefficient frontier can be defined, from the output perspective as, a convex hull defined by the minimum output level given an input level, for which it would not be rational to produce output levels less than the frontier value. Here “rational" implies the cost of the inputs is equal to the income generated from the outputs.

While the existence of an inefficient frontier is motivated by cost and price information, this data is not always available. Thus convex combinations of the most

inefficient observed units estimate the inefficient frontier, in a manner analogous to the definition of the efficient frontier. Similarly, from the input perspective the inefficient frontier is a convex hull defined by the maximum input level given an output level, for which it would not be rational to use input levels greater than the frontier value.

An observation may lie outside of the inefficient frontier if there is error in the measurement or entry of the data, if the observation is a chance instance of a low probability situation, or the observation may not truly belong to the group under evaluation. For any of these reasons a data point should be removed from the analysis. Further, a cross sectional analysis is a snapshot of a dynamic market. According to basic economic theory inefficient producers should be driven out of markets. Thus extremely inefficient observations in a particular cross-section could represent firms going out of business and in later cross sections would not be observed.

When outlier detection techniques are applied to the inefficient frontier, the envelopment concept is relaxed in order to quantify the degree to which each unit on the inefficient frontier (call these units completely inefficient units) is an outlier. Completely inefficient units are defined as units producing the lowest possible output level for a given input level or using the highest possible input level for a given output level of all units in the set being evaluated. These units represent in some sense the worst possible performance within the observed production possibility set.

When the inefficient frontier is included the production possibility set is defined as:

$$\hat{P} = \left\{ \begin{array}{l} (x, y) \mid y \le Y\lambda, x \ge X\lambda, i'\lambda = 1, \lambda \in R_+^n \\ \text{and } \; y \ge Y\mu, x \le X\mu, i'\mu = 1, \mu \in R_+^n \end{array} \right\} \tag{4.11}$$

For this new definition, a Shephard's input inefficient distance function can be defined:

$$D_{iII}(x_i, y_i) = \min\left\{\psi_{iI} \middle| {}^{x_i}\!/\!_{\psi_{iI}} \in L(y)\right\} \tag{4.12}$$

where the subscripts on D indicate unit, input and inefficiency, respectively. Similarly, a Shephard's output inefficient distance function can be defined as

$$D_{iOI}(x_i, y_i) = \max\left\{\psi_{iO} \middle| {}^{y_i}\!/\!_{\psi_{iO}} \in K(x)\right\} \tag{4.13}$$

The shape of the one-input, one-output inefficient frontier shown in figure 4.1 is an approximation of the true inefficient frontier, constructed from the observed data.

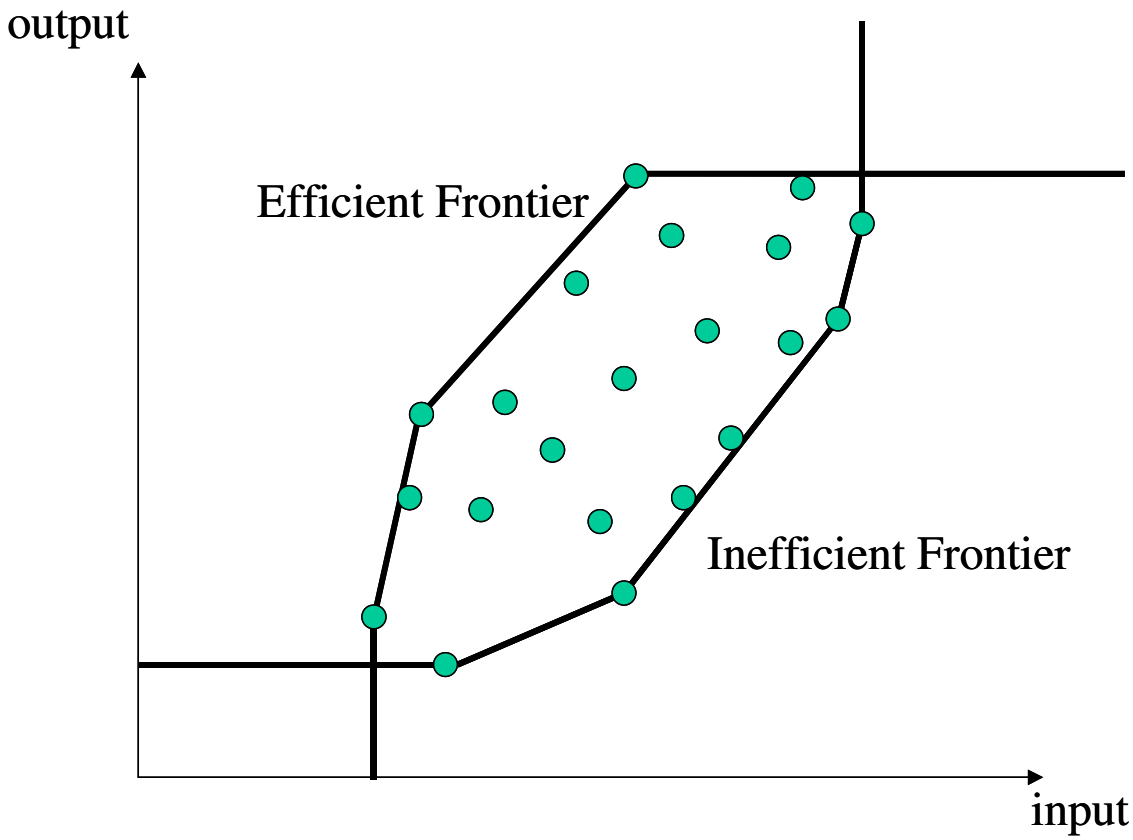

Figure 4.1: The inefficient and efficient frontiers for one input, one output

However, it can be argued under rather general assumptions about price, cost and risk aversion the true inefficient frontier should be convex. Assume a unit is small relative to the entire industry, thus its price is:

$$price = \alpha_0 \tag{4.14}$$

where α_0 is the constant market price. Further assume cost is made up of a fixed component and a variable component. Thus cost per unit is equal to

$$\text{cost per unit} = \beta_0 + \beta_1 / x \tag{4.15}$$

where x is the amount of input purchased, β_0 is the variable cost, and β_1 is the fixed cost, and all costs are positive. The condition that holds along the inefficient frontier is total cost is equal to income, implying this is the breakeven point. Below this level the cost would be greater than the income and a rational unit would go out of business. Thus substituting into the condition

$$\begin{gathered} \text{Total Cost} = \text{Income} \\ \text{cost per unit} * \text{input} = \text{price} * \text{output} \\ \left(\beta_0 + \beta_1 / x\right) * x = \alpha_0 * y \\ \beta_0 x + \beta_1 = \alpha_0 y \\ \left(\frac{\beta_0}{\alpha_0}\right) x + \left(\frac{\beta_1}{\alpha_0}\right) = y \end{gathered} \tag{4.16}$$

For this case the relationship between x and y along the inefficient frontier is described by a line with the slope $\left(\frac{\beta_0}{\alpha_0}\right)$ which is positive because both components are positive. If cost only has a variable component the shape of the frontier will not change.

However, if we further assume the firm is risk averse and the demand curve is downward sloping in price, we would expect the relationship between input and output to change, as more input is required. Now as input increases the rate of output growth

would need to increase, requiring more and more output per unit of input because of the firms risk aversion. However, at some point the cost of producing one more unit of output, beyond output level H, exceeds the benefit or the price for which that unit of output can be sold. At this point it would not be rational to invest any further in inputs as shown in Figure 4.2.

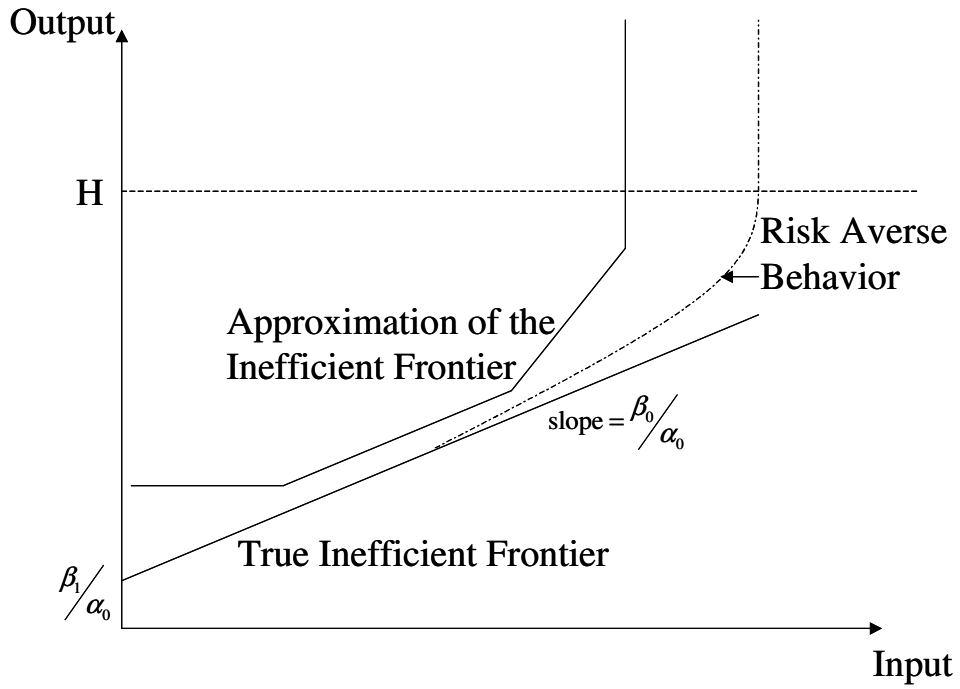

Figure 4.2: The true and approximate inefficient frontiers

Each observed unit is in operation thus must be operating above the true inefficient frontier. The approximation of the inefficient frontier is the minimum convex hull containing all the observed units. This is a conservative estimate for the true inefficient frontier.

4.3.1.1 The Single-Output Inefficient Production Frontiers

Figure 4.1 shows the inefficient frontier constructed for a single output and a single input. However, even simple production processes typically use several inputs to produce an output. The result of the process may be an individual good or this output may be an aggregation of several outputs. In this case we can define an input oriented inefficiency measure relative to an inefficient frontier as,

$$IE_{iII}\left(y_i, x_i\right) = \max\left\{\phi_{iII} : y_i \geq g\left(\phi_{iII} x_i\right)\right\} \tag{4.17}$$

Where g represents the least efficient production function. Here $IE_{iII}\left(y_i, x_i\right) \geq 1$ and equals one when the unit is completely inefficient. Similarly, if only a single output is produced, an output-oriented measure of inefficiency relative to an inefficient frontier is given by the function

$$IE_{iOI}\left(y_i, x_i\right) = \min\left\{\phi_{iOI} : \phi_{iOI} y_i \geq g\left(x_i\right)\right\} \tag{4.18}$$

where $IE_{iO}\left(y_i, x_i\right) \leq 1$ and equals one when the unit is completely inefficient.

4.3.1.2 The Multiple-Output Inefficient Production Frontiers

While it sometimes is possible to aggregate inputs, often there is no convenient weighting system that quantifies the relative values of the outputs. Thus often a multi-input / multi-output model is needed. The analytical framework is very similar to the single-output case; however, the single-output production frontier is replaced with Shephard's distance function. Input distance functions are used to define input-oriented measures of efficiency, and output distance functions are used to define output-oriented measures of efficiency. Thus if any number of outputs is produced, an input-oriented measure of inefficiency relative to an inefficient frontier is given by the function

$$IE_{iI}\left(y_i, x_i\right) = \max\left\{\phi_{iII} : \psi_{iII}\left(y_i, \phi_{iII} x_i\right) \geq 1\right)\right\} \tag{4.19}$$

The inefficient frontier with respect to the subset X(y) can be denoted as $\partial X_{in}(y)$ and found by

$$\partial X_{in}(y) = \left\{x \mid x \in L(y), \phi_{II} x \notin L(y) \forall\ 1 < \phi_I\right\} \tag{4.20}$$

Then the inefficiency estimate calculated from the input perspective can be found by solving the following linear program

$$\begin{aligned} &\max_{\phi_{iII}, \mu\lambda}(\phi_{iII}), \\ &s.t.\ \ -y_i + Y\mu \leq 0, \\ &\quad \phi_{iII} x_i - X\mu \leq 0, \\ &\quad \sum_{j=1}^{N} \mu_j = 1 \\ &\quad \mu_j \geq 0 \end{aligned} \tag{4.21}$$

For completeness, if any number of outputs is produced, an output-oriented measure of inefficiency relative to an inefficient frontier is given by the function

$$IE_{iO}\left(y_i, x_i\right) = \min\left\{\phi_{iOI} : \psi_{iOI}\left(\phi_{iOI} y_i, x_i\right) \leq 1\right)\right\} \tag{4.22}$$

The inefficient frontier with respect to the subset Y(x) can be denoted as $\partial Y_{in}(x)$ and found by

$$\partial Y_{in}(x) = \left\{y \mid y \in Y(x), \phi_{OI} y \notin Y(x) \forall\ 0 < \phi_{OI} < 1\right\} \tag{4.23}$$

We also show the linear program for calculating the inefficiency estimate from the output perspective

$$
\begin{aligned}
&\min_{\phi_{iOI},\mu}(\phi_{iOI}),\\
&s.t.\ \ -\phi_{iOI}\, y_i + Y\mu \le 0,\\
&\qquad x_i - X\mu \le 0,\\
&\qquad \sum_{j=1}^{N} \mu_j = 1\\
&\qquad \mu_j \ge 0
\end{aligned}
\tag{4.24}
$$

With these concepts and terminology defined, we can now explain how to use the leave-one-out outlier detection method of Wilson [1995] relative to an inefficient frontier.

4.3.2 Outlier Detection Relative to the Efficient and Inefficient Frontiers

One outlier detection method suggested by Wilson [1995] calculates the leave-one-out efficiency estimate to give a measure of the degree to which an observation is an outlier. While Wilson only searches for outliers relative to either an input or an output orientation, Simar [2003] suggests an observations should be distant from both an input and an output orientation in order to be an outlier. For identifying outliers relative to an efficient frontier we will heed Simar's suggestions and require the observation to be distant from both perspectives. To quantify distant, a threshold value needs to be selected. If a threshold value is chosen for one of the orientations, the reciprocal value should be used for the other orientation to specify symmetrical thresholds.

Relative to an inefficient frontier, if an observation is found to be both below this threshold value for input oriented analysis and above the reciprocal value for output oriented analysis then the observation will be flagged as an outlier requiring further inspection. A value of 1.5 for the input oriented estimate threshold, means an observation being evaluated is excluded if the reference point on the inefficient frontier requires less than $2/3$ the input. Similarly the reciprocal value, 0.66, means an

observation being evaluated is excluded if it produces more than $2/3$ the output of the reference pint on the inefficient frontier.

This is an example of a weak outlier threshold criterion. Of course more rigorous criteria could be selected by picking a larger value for the input oriented estimate threshold or by selecting a smaller value for the output oriented estimate threshold. Wilson does not provide any guidance in the selection of these threshold criteria for the efficient frontier and Simar [2003] states that threshold values will be closely related to the data generation process which is specific for each group evaluated. Thus this value should be selected on a case-by-case basis.

The leave-one-out input oriented DEA inefficiency estimate is the distance from the inefficient observation to the inefficient frontier of the data set, not including the observation under evaluation, and can be computed using the following linear program:

$$\begin{aligned} &\max_{\phi^*_{iII},\mu^*_i}(\phi^*_{iII}), \\ &s.t.\ \ -y_i + Y^{(i)}\mu^*_i \le 0, \\ &\quad \phi^*_{iII}x_i - X^{(i)}\mu^*_i \le 0, \\ &\quad \sum_{j=1}^{N}\mu^*_j = 1 \\ &\quad \mu^*_j \ge 0 \end{aligned} \tag{4.25}$$

In (4.25), ϕ^*_{Ii} is the input-oriented inefficiency estimate for the ith unit, μ^*_i is a vector of intensity variables, $X^{(i)} = \left[x_j\right] \forall j \neq i$, $Y^{(i)} = \left[y_j\right] \forall j \neq i$, $x \in R^p_+$ denotes a $(1 \times p)$ vector of inputs, $y \in R^q_+$ denotes a $(1 \times q)$ vector of outputs, and N is the number of observations. The variables $X^{(i)}$ and $Y^{(i)}$ have dimensions $\left(p \times (N-1)\right)$ and $\left(q \times (N-1)\right)$, respectively, and μ^*_i has dimensions $\left(1 \times (N-1)\right)$. Similarly, the leave-

one-out output oriented DEA inefficiency estimate can be calculated by the linear program

$$\begin{aligned} &\min_{\phi^*_{iOI},\mu_i^*}(\phi^*_{iOI}), \\ &s.t.\ \ -\phi^*_{iOI}y_i + Y^{(i)}\mu_i^* \le 0, \\ &\qquad x_i - X^{(i)}\mu_i^* \le 0, \\ &\qquad \sum_{j=1}^{N}\mu_j^* = 1 \\ &\qquad \mu_j^* \ge 0 \end{aligned} \tag{4.26}$$

Both (4.25) and (4.26) must be solved one time for each observation in order to develop a set of leave-one-out inefficiency estimates for all observations. Observations that are candidates for outliers will have leave-one-out input oriented DEA inefficiency estimates, ϕ^*_{iII}, less than one or leave-one-out output oriented DEA inefficiency estimates, ϕ^*_{iOI}, of greater than one. However, because not all completely inefficient observations should be considered outliers a threshold value must be established.

A common problem facing outlier detection methods is the masking effect. Rousseeuw and van Zomeren [1990] give a detailed discussion of this problem; in essence, the presence of an outlier hides or masks the presence of another outlier. The leave-one-out method is based on a nearest neighbor type criterion, and is particularly vulnerable to this effect. A method suggested by Simar [2003] and Wilson [1995] to ameliorate this problem is to apply an outlier detection process in an iterative fashion, i.e., the outlier detection method should be applied, outliers identified and removed, and the method applied again on the smaller set. This process could be applied a specified number of times or until the number of outliers identified in an iteration is below a specified level. However, if for example there were units coming from two data

generation processes, we could expect that the data would fall into two clusters and the typical outlier detection methods would not necessary identify this phenomena. This is why some have recommended cluster analysis should be the first step of outlier analysis (see for example footnote 3 in Simar 2003); we assume the problem of mixed populations is not present, or has been dealt with already through an appropriate clustering method.

4.4 Inefficient Frontier: Practical Implementation

In this section we will demonstrate the use of outlier detection methodology considering both efficient and inefficient frontiers. It will be shown that simply identifying and removing outliers relative to an efficient frontier, and ignoring the inefficient frontier, can significantly impact the conclusions drawn from the results of the second stage of a two-stage semi-parametric model.

Here we use the classic Banker and Morey [1986b] data set for pharmacies in the state of Iowa. There are 69 observations, 2 outputs, 3 inputs, and 1 environmental variable. The environmental variable is population and the continuous values for population are used (rather than the categorical variable constructed by Banker and Morey). For more information about the data set, see Banker and Morey [1986b].

To begin, a critical value for outlier detection should be specified. The rather strict value of 1.1 was selected for the efficient frontier input oriented evaluation and the inefficient frontier output oriented evaluation. The reciprocal value of 0.91 was used for the efficient frontier output oriented evaluation and the inefficient frontier input oriented evaluation. Because the iterative method was used, the iteration on which an observation was identified as an outlier is also noted in Table 4.1.

Table 4.1: Pharmacy ID number and the iteration for outlier test

	Efficient Frontier	Inefficient Frontier
1st Iteration	5 69	17 69
2nd Iteration	6 41 55	15 23 46
3rd Iteration	17	47
4th Iteration	4 7 12 44 53 65 67	

Note that observation 69 was flagged both for the efficient frontier outlier analysis and the inefficient frontier outlier analysis. An observation which is consuming a small amount of an input and producing the small amount of an output could be flagged both by the efficient and inefficient analysis. Similarly, an observation which is consuming the large amounts of an input and producing the large amounts of an output could be flagged by both analyses. Observation 69 is a small pharmacy.

The two-stage bootstrapping method, algorithm 2 in Simar and Wilson [2005] was used to estimate the equation

$$\delta_i = \beta_0 + z_i\beta + \varepsilon_i \tag{4.27}$$

where z is a (69×1) vector of the population values and δ_i is the input efficiency of unit i. The 95% bootstrap confidence interval for the parameter β based on 56 points remaining in the data set after removing outliers relative to the efficient frontier was [-1.982, 1.712]. Thus the result of the analysis would conclude the population has no effect on the efficiency of a pharmacy in Iowa. However, if the bootstrapping method is

used on the 52 point data set with outliers removed based on both the efficient and inefficient frontier, the bootstrap confidence interval at the 95% level is [-0.3460, -0.0005]. This result indicates that efficiency is inversely related to population of the area in which the pharmacy is located. The main point of this example is to show that misleading conclusions can be drawn from the second stage analysis if outliers are not identified and treated for both the efficient and inefficient frontiers.

4.5 Conclusion

This chapter describes an outlier detection methodology, and introduces the inefficient frontier. The inefficient frontier's value as a concept to aid in outlier detection is demonstrated. Further this chapter implements the iterative outlier detection method previously discussed in both Simar [2003] and Wilson [1995] and demonstrates the Simar and Wilson [2005] two-stage semi-parametric method for the Banker and Morey [1986b] data set with outliers removed based only on the efficient frontier and for the data set with outliers identified based on both the efficient and inefficient frontiers. It is shown that the conclusions drawn based on the results of the two different data sets can be different and the use of outlier detection based on both the efficient and inefficient frontiers is recommended.

CHAPTER 5

THE HYPERBOLIC ORIENTED EFFICIENCY MEASURE AS A REMEDY TO INFEASIBILITY OF SUPER EFFICIENCY MODELS

5.1 Introduction

Performance measurement is an important issue in any enterprise. The ability to distinguish between top performing units and poorer performing units is important for a variety of reasons. In a public environment or in a large company, performance measurement allows resources to be allocated to the units which are the most productive. In a competitive environment it allows poor performers to understand the quality of their performance and to apply benchmarking techniques to guide them toward improvement. However, many industries or units operate in multi-input multi-output environment. To understand performance, the set of relevant inputs and outputs need to be considered simultaneously.

The study of performance measurement has an important starting point in T.C. Koopmans [1951] book where he developed his definition of efficiency.

> "A possible point in the commodity space is called efficient whenever an increase in one of its coordinates (the net output of one good) can be achieved only at the cost of a decrease in some other coordinate (the net output of another good)."

The term commodity space is more commonly referred to as the production possibility set, meaning a set of all points, representing input (vector) and output (vector) pairs such that the input can be used to produce the output. Thus a technically inefficient producer

could, by improving its performance, produce its output with less of at least one input, or could use its inputs to produce more of at least one output.

Based on these ideas, Shephard [1953] developed the input distance function in which he finds the equaproportional reduction of inputs for which the production of a given output set is still feasible. Later Shephard [1970] developed the output distance function in which input levels are fixed and the equaproportional increase of outputs is identified. These two measures represent two common measures of efficiency. Farrell [1957] introduced an efficiency measurement independent of Shephard, which was later shown to be the dual of Shephard's distance functions and is the basis for the technique referred to as Data Envelopment Analysis (DEA) Charnes, Cooper and Rhodes [1978]. The decision to take an input or an output orientation when using the distance function or DEA has been a critical decision typically made based on an argument of whether the enterprise is cost minimizing (input orientation) or profit maximizing (output orientation) (see, e.g., Fare and Primont [1995]).

Fare, Grosskopf and Lovell [1985] introduced the hyperbolic distance function as an alternative to selecting either an input orientation or an output orientation. This measure is a simultaneous equiproportionate expansion of output and contraction of inputs. Because it requires solving a nonlinear program, rather than the linear programs associated with Shephard distance functions and data envelopment analysis, this measure has been slow to gain popularity. Research in the area has been limited. Fare, Grosskopf and Lovell [1994] further explored the topic, proving a variety of properties of the hyperbolic distance function. Fare, Grosskopf and Zaim [2002] related the measure to return to the dollar. Cuesta and Zofio [2005] used the hyperbolic distance function in a

parametric context to extend a translog production function to a multi-input/multi-output production environment.

As with Shephard's distance function, hyperbolic distance functions also can be applied in super efficiency models to measure the efficiency of an observation outside the production possibility set. Two important applications of super efficiency models are outlier detection and calculating the Malmquist index.

Johnson and McGinnis [2005a] developed an outlier detection model which calculates both the input oriented and the output oriented super efficiency measures. An observation is flagged for further investigation if both the input oriented and the output oriented measures exceed a specified critical value. However, because these measures are calculated using linear programming and can return infeasible solutions (as discussed below), an observation may be identified as an outlier based on only one orientation or might be flagged because both orientations return infeasible solutions. Thus the infeasibility of the input or output orientation causes the outlier detection method to make decisions based on incomplete information.

As noted by Zofio and Lovell [2001], using the hyperbolic distance measure allowed them to calculate Malmquist Productivity Index for their data set. If it could be shown the hyperbolic oriented measure could always be calculated, particularly in the cases when the Shepard's input oriented or the output oriented distance measures could not, then the hyperbolic oriented measure would provide analysts with a more reliable method to calculate the Malmquist Productivity Index.

A hyperbolic oriented super efficiency measure can be developed by applying the hyperbolic measure to evaluate the efficiency of an observation relative to a reference set

which does not include the observation under evaluation. This chapter will show that hyperbolic oriented super efficiency models have advantages over standard super efficiency models for two reasons. First, a sufficient condition for feasibility is for all observation data to be positive. Second, while it is still possible to have infeasible solutions when zeros are allowed in the data domain, the conditions for infeasibility are more limited for the hyperbolic measure than for standard DEA-based super efficiency models.

The chapter is structured as follows: section 5.2 provides descriptions of the standard super efficiency models and the super efficiency hyperbolic model. Section 5.3 shows feasible solutions are possible for super efficiency hyperbolic model in cases when standard super efficiency models cannot find feasible solutions. Concluding remarks are given in section 5.4.

5.2 Standard Super Efficiency Models and a Super Efficiency Hyperbolic Model

Super efficiency models measure the efficiency of an observation outside the production possibility set. These models are a special case of data envelopment analysis (DEA) models. Typically input oriented DEA efficiency estimates are on the range of $(0,1]$ and output oriented DEA efficiency estimates are on the range of $[1,\infty)$. However, for super efficiency models the range of $(0,\infty)$ is possible for either orientation. The super efficiency model was first referenced in Banker, Das and Datar [1989] as an outlier detection method developed in a separate working paper by the same lead author. The entire model later appeared in Anderson and Petersen [1993] as a method for developing a full ranking of observations.

Since then, the method has been used in a variety of situations. Charnes, Rousseau and Semple [1996] and Zhu [1996] used the method to study the sensitivity of the efficiency classification (see also Seiford and Zhu [1998]). Fare, Grosskopf and Lovell [1994] employed these models to measure productivity and technology change. Thrall [1996] used them to identify extreme efficient observations, and Wilson [1995] and Johnson and McGinnis [2005a] imbed the model in computational methodologies to find outliers.

It has been noted that under various conditions the standard super efficiency model, taking an input or an output orientation relative to a variable returns to scale frontier, may not be solvable and is said to have an infeasible solution. This has been noted by Thrall [1996] and Zhu [1996] elaborated by identifying certain zero patterns appearing in the data which cause infeasibility of the super efficiency model with or without the returns to scale assumption. Seiford and Zhu [1999] provided the most comprehensive discussion of this topic, defining exhaustively the conditions under which either an output or an input oriented super efficiency model would not have a feasible solution. In contrast to Zhu [1996], Seiford and Zhu [1999] only studied the infeasibility of the super efficiency models, where they assume that all input and output observation data are positively-valued.

Given n observations, and associated with each observation ($j = 1, 2..., n$) a vector of inputs, x_j, and a vector of outputs y_j, let x_{ij} be the ith input in the set P of inputs and let y_{ij} be the ith output in the set Q of outputs. Under the assumption of variable returns to scale two super efficiency DEA models can be expressed as shown in figure 1.

Input-Oriented	Output-Oriented
$\min \theta$	$\min \Theta$
$s.t. \quad \sum_{j=1}^{n} \lambda_j x_j \le \theta x_0$	$s.t. \quad \sum_{j=1}^{n} \lambda_j x_j \le x_0$
$\sum_{j=1}^{n} \lambda_j y_j \ge y_0$	$\sum_{j=1}^{n} \lambda_j y_j \ge \Theta y_0$
$\sum_{j=1}^{n} \lambda_j = 1$	$\sum_{j=1}^{n} \lambda_j = 1$
$\theta, \lambda_j \ge 0, \; j \ne 0$	$\Theta, \lambda_j \ge 0, \; j \ne 0$

Figure 5.1 Standard Super Efficiency Models

where (x_0, y_0) represents the observation under evaluation. Note that the super efficiency model differs from a standard variable returns to scale (VRS) DEA model in that the observation under evaluation is excluded from the reference set.

Using the same notation, a super efficiency hyperbolic model can be expressed as

$$\begin{aligned} &\min \theta \\ s.t. \quad &\sum_{j=1}^{n} \lambda_j x_j \le \theta x_0 \qquad (\text{Set of Input Constraints}) \\ &\sum_{j=1}^{n} \lambda_j y_j \ge {}^{1}\!/_{\theta}\, y_0 \quad (\text{Set of Output Constraints}) \\ &\sum_{j=1}^{n} \lambda_j = 1 \qquad (\text{Convexity Constraint}) \\ &\theta > 0, \; \lambda_j \ge 0, \; j \ne 0 \end{aligned} \tag{5.1}$$

The infeasibility problem in two dimensions is illustrated in figure 5.2. Using a standard DEA model, the efficient frontier is constructed using points A, B, and C. The frontier used to measure the super efficiency of A uses only B and C to construct the frontier. The super efficiency for B can be calculated by either orientation; however, the super

efficiency can only be calculated from an input orientation for A and from an output orientation for C.

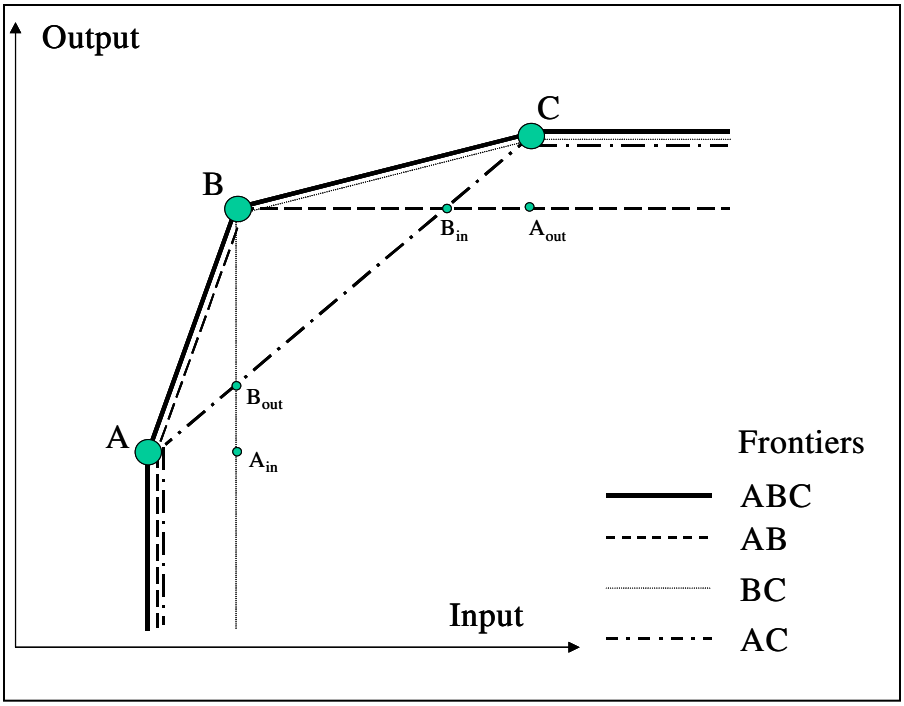

Figure 5.2 Super efficiency illustrated in two dimensions

To better understand the cause of an infeasible result for a DEA model we can examine the linear program. Looking first at the input oriented super efficiency model we can see there are three types of constraints, those related to the inputs, to the outputs, and to convexity. Each output constraint has the form

$$\sum_{j=1}^{n} \lambda_j y_j \geq y_0 \tag{5.2}$$

In the variable returns to scale model the λ_j are non-negative and sum to one, which implies all λ_j values are less than or equal to one. Thus if y_0 is greater than all y_j values for any of the outputs in the reference set, the constraint associated with that

output cannot be satisfied and the problem is infeasible. If $y_j \geq y_0$ for all outputs for a given j then a solution to the input oriented super efficiency model can always be found. A more general result is given by Theorem SZ7 in Seiford and Zhu [1999], which states

> Theorem SZ7
> The input oriented super efficiency variable returns to scale model is infeasible if and only if $g^* < 1$, where g^* is the optimal value to the following linear program

$$\begin{aligned} g^* &= \max g \\ s.t.\ \sum_{j=1}^{n} \lambda_j y_j &\geq g\, y_0 \\ \sum_{j=1}^{n} \lambda_j &= 1 \\ \lambda_j &\geq 0,\ j \neq 0 \end{aligned} \tag{5.3}$$

Notice there is no such problem with the input constraints in an input oriented problem. Assume a vector of λ_j can be found to satisfy the output constraints. Use these λ_j in the input constraints. Since it was assumed all data are positive, a vector of λ_j implies a particular convex combination of these positive values which must itself be a positive value. Thus each input constraint implies

$$\text{Positive Value} \leq \theta x_0 \tag{5.4}$$

Where x_0 is also a positive number, thus a value of θ to satisfy this constraint can always be found and the maximum such value over all input constraints will satisfy all the input constraints.

Similarly for the output oriented super efficiency model the infeasibility problem arises from an inability to satisfy the input constraints.

$$\sum_{j=1}^{n} \lambda_j x_j \leq x_0 \tag{5.5}$$

If x_0 is less than x_j for all j for any input, then the constraint associated with that input cannot be satisfied. If $x_j \le x_0$ for all inputs for a given j then a solution to the output oriented super efficiency model can always be found. A more general result is given by Theorem SZ2 in Seiford and Zhu [1999].

Theorem SZ2
The output oriented super efficiency model relative to variable returns to scale frontier is infeasible if and only if $h^* > 1$, where h^* is the optimal value to the following linear program

$$\begin{aligned} h^* &= \min h \\ s.t.\ & \sum_{j=1}^{n} \lambda_j x_j \le h\, x_0 \\ & \sum_{j=1}^{n} \lambda_j = 1 \\ & \lambda_j \ge 0,\ j \ne 0 \end{aligned} \tag{5.6}$$

At this point it is valuable to note that $h^* > 1$ in theorem 2 and $g^* < 1$ in theorem 7 can both be true simultaneously. This is because the linear program in theorem SZ2 only involves inputs and the linear program in theorem SZ7 only involves outputs thus the results from the two linear programs are not related and would be determined independently. This situation in two dimension is illustrated in figure 5.3.

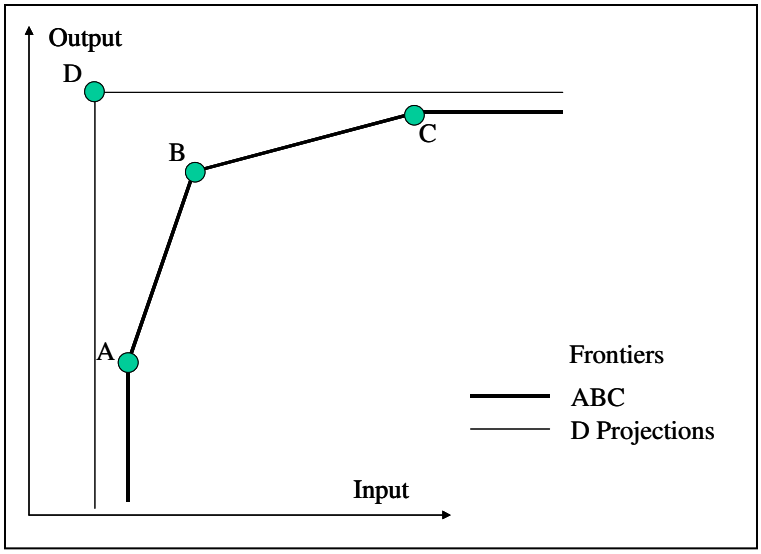

Figure 5.3 Super efficiency illustrated in two dimensions when the input and the output oriented measures are both infeasible

5.3 Feasibility of Hyperbolic Efficiency Measure

In Theorems SZ2 and SZ7, Seiford and Zhu [1999] recognized that infeasibility occurs when the input constraints in an output oriented model or the output constraints in an input oriented model are not satisfiable. We now will show for the hyperbolic orientation that both sets of constraints are always satisfiable and thus infeasibility is not possible when the hyperbolic orientation is used.

Theorem 1

When all input and output values are positive, the hyperbolic oriented super efficiency model under a variable returns to scale production frontier always has a feasible solution.

Proof.

For any λ satisfying the convexity constraint, the input constraints can be satisfied by selecting $\theta \geq U(P)$ where

$$U(P) = \max_{i \in P} \left(\left(\max_{j=1...n} x_{ij} \right) \Big/ x_{i0} \right) \tag{5.7}$$

Clearly, for any i, $U(P)x_{i0}$ is larger than x_{ij}, for all j, thus larger than any convex combination of the x_{ij}. By a similar argument, for any λ satisfying the convexity constraint, the output constraints can be satisfied by selecting $\theta \geq 1/L(Q)$ where

$$L(Q) = \min_{i \in Q} \left(\min_{j=1...n} y_{ij} / y_{i0} \right) \tag{5.8}$$

It is possible to find a θ satisfying both conditions simultaneously, from

$$\theta = \max \left(U(P),\ 1/L(Q) \right) \tag{5.9}$$

This means a feasible value of θ always exists □

As long as all input and output values are positive, a hyperbolic oriented super efficiency model always has a feasible solution. Even if this assumption is relaxed, the hyperbolic oriented super efficiency model may have a solution.

Relaxing the assumption that all data are positive, theorem 2 states the conditions under which hyperbolic oriented super efficiency model has a feasible solution. Note the output constraints in the hyperbolic oriented super efficiency model are satisfied for any vector of λ_j, even when some y_{i0} is zero. Thus zero values for outputs do not lead to infeasibility, and only zero values of inputs need be addressed.

Theorem 2

Let $P' \subseteq P$, $\ni$ $\forall i \in P'$, $x_{i0} = 0$. A necessary and sufficient condition for the hyperbolic oriented super efficiency model to have a feasible solution is that there is at least one observation, say x_s, in the reference set, such that $x_{is} = 0$, $\forall i \in P'$

Proof

Sufficiency: A feasible solution can be found by setting $\lambda_s = 1$ to satisfy the input constraints corresponding to P'; and applying theorem 1 to $P \setminus P'$ to determine a sufficiently large θ.

Necessity: Suppose there is a feasible solution and $\nexists x_s \ni x_{is} = 0 \ \forall i \in P'$, i.e., $\forall x_s$, $\exists i \in P' \ni x_{is} > 0$. In order to satisfy the ith constraint, we must have $\lambda_s = 0$. However, if $\lambda_s = 0$ for all s, then $\sum_j \lambda_j = 1$ cannot be satisfied. Thus the initial assumption is contradicted, and if there is a feasible solution, there must be at least one $x_s \ni x_{is} = 0 \ \forall i \in P'$ □

We have shown that the hyperbolic oriented super efficiency measure can be calculated in cases when the input or the output oriented super efficiency measure cannot. Further we have shown that as long as a single observation in the reference set has zeros for the same input set as the observation under evaluation, the hyperbolic oriented super efficiency measure is computable. Finally zeros in the output space do not affect the feasibility of the hyperbolic oriented super efficiency measure.

5.4 Conclusion

This chapter studies the use of the hyperbolic oriented super efficiency measure and its benefits relative to the more traditional input or output oriented super efficiency measures. The hyperbolic efficiency measure has been slow to gain popularity in part because of its increased computational burden. It requires solving a nonlinear program rather than a linear program. However for a data set of 390 observations, we have found

it takes 8.97 seconds to calculate the input efficiency scores and 15.6 seconds to calculate the hyperbolic oriented efficiency estimates for all observations. The envelopment formulation, using a 2.3 GHz Sun System running GAMS software with a CPlex solver, was used for the calculations.

Super efficiency has been used for sensitivity analysis, productivity and technology change (such as the Malmquist index), ranking DEA efficient observations and for outlier detection. The benefits of using the hyperbolic oriented super efficiency measure could be realized for each of these applications.

Our results indicate it is still possible to have infeasible hyperbolic oriented measures when the observation under consideration has values of zero for a set of inputs for which no observation in the reference set has zeros for the same set inputs. However, this is a rather weak condition and if it is not satisfied the analyst may question if all observations under evaluation are truly using the same production technology.

The hyperbolic oriented efficiency measure requires a minimal increase in computational time and has two benefits: it can be calculated for cases when the input or output oriented measures can not be calculated, and it allows the comparison of a broader group of observations by allowing zeros as input values. These results make the hyperbolic oriented efficiency measure a desirable option.

CHAPTER 6

A QUANTILE-BASED APPROACH FOR RELATIVE EFFICIENCY MEASUREMENT WITH MULTIPLE INPUTS AND OUTPUTS

6.1 Introduction

Production processes can be modeled in a variety of ways. Economists typically employ a cost or production function to characterize the relationship between inputs, outputs, and costs. However, nonparametric methods beginning with Farrell [1957] have also gained popularity. These methods measure efficiency relative to a comparison set of firms. Shephard [1970] introduced distance functions to characterize the relationship between observed production levels and a production frontier. Boles [1971] developed a method to measure efficiency which implemented Farrell's ideas and was later named data envelopment analysis (DEA) by Charnes, Cooper and Rhodes [1978]. The primary competing methodology is one developed in Aigner, Lovell and Schmidt [1977], the stochastic frontier approach (SFA). This method makes assumptions about the distribution of efficiency estimates in order to separate efficiency from the error term in the residual of a standard linear regression. Each method has advantages and disadvantages; supporters of DEA often cite the lack of restrictive distributional assumptions as their primary advantage while SFA supporters cite the ability of SFA to consider measurement and specification errors as a key advantage.

Griffin and Kvam [1999] introduced a quantile-based approach (QBA) for relative efficiency measurement. The advantages of their method are that it relies less on the

stochastic modeling than SFA methods, and it accounts for the firm's relationship to the other firms in the comparison set by acknowledging the firm's influence on an empirical model. This chapter will demonstrate how to use the QBA method in an environment with multiple inputs and multiple outputs.

Because actual efficiency levels are unobservable it is not possible to determine definitively which efficiency method is the most accurate. Thus the practioner is still left to weigh the advantages and disadvantages of each method when deciding how best to measure efficiency for their particular problem. However, the multi-input/multi-output quantile-based analysis (MQBA) developed in this chapter no longer has the disadvantage of only being able to handle one output or using an arbitrary aggregation scheme to combine several outputs.

The layout of this chapter is as follows. In Section 6.2, we describe the mathematical background necessary to develop a simultaneous multi-output / multi-input model. The DEA and SFA models will be briefly summarized in Sections 6.3 and 6.4. MQBA will then be described and the calculations outlined in Section 6.5. A sample data set will be used in Section 6.6 to compare DEA, SFA and MQBA will be demonstrated on a sample data set. Conclusions will be presented in Section 6.7.

6.2 Mathematical Background

Shephard's distance function can be applied with either an input orientation or an output orientation. Here we will present the input orientation; however, for a description of output orientation see Coelli, Rao and Battese [1998] section 3.4.2. Define an input vector, $x \in R_+^P$ where P is the number of different inputs. The input set, $L(y)$, is all input quantity combinations that can achieve an output quantity y, where $y \in R_+^Q$ is the

set of all output vectors that can be produced. Similarly, Q is the number different outputs. Therefore,

$$L(y) = \{x \in R_+^N : x \text{ can produce } y\} \tag{6.1}$$

Shephard's input distance function can be defined over the input set $L(y)$ as

$$D_I(x, y) = \max\{\delta_I : (x/\delta_I) \in L(y)\} \tag{6.2}$$

For all x that are elements of the feasible input set, $L(y)$, the distance function, $D_I(x, y)$, will be greater than or equal to one.

As first shown by Lovell et al. [1993] Shephard's distance function can be used to aggregate inputs and to develop a constructed variable which carries information about all inputs. This constructed variable can then be used as a dependent variable for regression analysis, thereby allowing multiple inputs and outputs to be used at the same time. Here we will show the multi-output and multi-input production function described with a translog function. The translog input distance function is

$$\begin{aligned} \ln D_{Ii} = {} & \alpha_0 + \sum_{m=1}^{M} \alpha_m \ln y_{mi} + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} \alpha_{mn} \ln y_{mi} \ln y_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} \\ & + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{ki} + \sum_{k=1}^{K}\sum_{m=1}^{M} \sigma_{km} \ln x_{ki} \ln y_{mi}, i = 1, 2, ..., N \end{aligned} \tag{6.3}$$

Where D_{Ii} is the input distance function value for observation i. The necessary conditions for homogeneity of degree 1 are

$$\begin{aligned} & \sum_{k=1}^{K} \beta_k = 1, \\ & \sum_{k=1}^{K} \beta_{kl} = 0,\ k = 1, 2, ..., K, \\ & \sum_{k=1}^{K} \sigma_{km} = 0,\ m = 1, 2, ..., M \end{aligned} \tag{6.4}$$

Additionally, the restrictions to maintain symmetry are

$$\begin{aligned} &\alpha_{mn} = \alpha_{nm}, \;\; m,n = 1,2,...,M \\ &\beta_{kl} = \beta_{lk}, \;\; k,l = 1,2,...,K \end{aligned} \tag{6.5}$$

Finally the constraint to impose linear homogeneity in outputs is

$$\sum_{m=1}^{M} \alpha_m = -1 \tag{6.6}$$

The requirement of homogeneity implies

$$D_I(\omega x, y) = \omega D_I(x, y), \text{ for any } \omega > 0 \tag{6.7}$$

Thus, by selecting an arbitrary input, say the K[th] input, and setting $\omega = 1/x_K$,

$$D_I\left(x/x_M, y \right) = D_I(x, y)/x_M \tag{6.8}$$

With this, the translog function can be rewritten as

$$\begin{aligned} &\ln\left(D_{Ii}/x_{Mi} \right) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \ln y_{mi} + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} \alpha_{mn} \ln y_{mi} \ln y_{ni} + \sum_{k=1}^{K-1} \beta_k \ln x_{ki}^* \\ &+ \frac{1}{2}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} \beta_{kl} \ln x_{ki}^* \ln x_{ki}^* + \sum_{k=1}^{K-1}\sum_{m=1}^{M} \sigma_{km} \ln x_{ki}^* \ln y_{mi}, i = 1,2,...,N \end{aligned} \tag{6.9}$$

where $x_k^* = x_k/x_K$.

If we allow TL to represent the translog functional form then the above can be written more concisely as

$$-\ln(x_{Ki}) = TL\left(x_k/x_K, y_i, \alpha, \beta, \sigma \right) - \ln(D_{Ii}), \;\; i = 1,2,...,N \tag{6.10}$$

This equation can now be used by either SFA or QBA to calculate efficiency scores. The distance function and the constructed variable are the key to allowing multi-inputs and multi-outputs simultaneously.

6.3 Data Envelopment Analysis

Data Envelopment Analysis (DEA) is one of several models referred to as non-parametric and deterministic. What is meant by *deterministic* is the assumption of no error in the measurement of data. However, since efficiency is calculated based on an unknown frontier which is estimated by sample data, efficiency scores still have statistical properties as noted by Simar and Wilson [2000]. Among non-parametric methods, when convexity and free disposability are assumed, the calculation is referred to as data envelopment analysis (DEA). Many models are available for estimating efficiency scores; however, for our purposes we will focus on the variable returns to scale DEA model to allow the fairest comparison to QBA and SFA. The DEA production set can be written as

$$\hat{P} = \left\{ (x, y) \mid y \le Y\lambda,\ x \ge X\lambda,\ i\lambda = 1,\ \lambda \in R_+^n \right\} \tag{6.11}$$

where $\hat{P}$ is an estimate based on the observed pairs (x_i, y_i) of the actual production set P, $x \in R_+^p$ denotes a $(1 \times p)$ vector of inputs, $y \in R_+^q$ denotes a $(1 \times q)$ vector of outputs, n is the number of observations, $Y = [y_1 ... y_n]$, $X = [x_1 ... x_n]$, i is a $(1 \times n)$ vector of ones and λ is an $(n \times 1)$ vector of intensity variables. For a further description of the relationship between the input and output spaces see either Lovell [1994] or Charnes et al. [1993]. The linear program to calculate the efficiency scores in the input requirement space is

$$\begin{aligned} &\max_{u,v} (u' y_i + u_*), \\ &s.t.\ \ v' x_i = 1, \\ &\qquad u' y_j + u_* \le v' x_j,\ j = 1, 2, ..., N, \\ &\qquad u, v \ge 0 \end{aligned} \tag{6.12}$$

This linear program needs to be solved once for each firm, $i = 1,...,n$.

The variables u and v are often referred to as weights. These variables represent a relative value system or a weight assigned to each input or output. Because the linear program is solved once for each firm, the weights can be selected as to maximize the efficiency of the firm. This allows a firm to assign a zero value to inputs that are being over used or outputs that are being under produced.

A problem in DEA is that for small data sets often a large percentage of the observations are found to be efficient. If the number of inputs and the outputs are reduced this can help mitigate this problem. However, if the number of inputs or outputs is reduced to one, the multi-input / multi-output nature of the analysis has been lost.

Additional constraints can be added to the linear program to enforce a minimum weighting, which can also reduce the number of efficient observations. The formulation above is input oriented, thus the minimum output weight constraints are of the form:

$$u_s y_s \Big/ \sum_i u_i y_i \geq \beta_s, \quad s = 1,...,q \tag{6.13}$$

where β_s is a minimum output weight for output s. The minimum input weight constraints are of the form:

$$v_t x_t \geq \gamma_t, \quad t = 1,...,p \tag{6.14}$$

where γ_t is a minimum input weight for input t. The product $v\,x$ is already constrained to equal one in the linear program. Minimum weight constraints restrict the production possibility set, thus forcing inputs or outputs to contribute to the efficiency estimate calculation. DEA model with weight restrictions will be referred to as DEA-W.

6.4 Stochastic Frontier Approach

Stochastic frontier approach (SFA) was developed to calculate efficiency while allowing for random error. Neither random error nor inefficiency can be observed directly, so separating them requires an assumption. SFA employs a *composed error* model in which inefficiencies are assumed to follow an asymmetric distribution, usually the half-normal, while random errors are assumed to follow a symmetric distribution, usually the standard normal Aigner, Lovell and Schmidt [1977].

The independent variables for the SFA regression model are the inputs $x = [x_1, ..., x_P]$, and the dependent variable is the output y. Thus the standard SFA equation is

$$y_i = f(x_i; \beta) + \varepsilon_i, \ i = 1, ..., N \tag{6.15}$$

The original SFA models only allowed a single-output. However, using the Shephard's distance function and the constructed variable, it is possible to estimate a model with multiple outputs as described above. The error term is given by $\varepsilon = \mu + \nu$, where $\mu \geq 0$ represents inefficiency and ν represents random noise.

Jondrow et al. [1982] developed a means of finding the individual efficiencies for each data point. Using the conditional distribution of the asymmetric error given the whole error, the two components can be separated. The technical inefficiency is given by $\hat{E}[u_i \mid v_i + u_i]$ and estimators for statistical noise are derived residually by means of

$$\hat{E}[v_i \mid v_i + u_i] = y_i - x_i\hat{\beta} - \hat{E}[u_i \mid v_i + u_i], \quad i = 1, ..., N \tag{6.16}$$

which provides conditional estimators for v_i. While the shape of the distribution of the inefficiency term has been the focus of much debate, SFA has the desirable property that regardless of the distribution imposed, the ordering of DMUs remains nearly the same.

6.5 Quantile-based approach

John Tukey developed the studentized deleted residual in Tukey [1958], sometimes called the jackknife residual, based upon Quenouille [1956]. The deleted residual, the base of all quantile statistics, is calculated via a regression equation; however, the parameters are estimated based on $n-1$ observations, excluding the observation under evaluation. The estimated parameters are then multiplied by the independent variable values and the result is subtracted from the dependent variable value of the observation under evaluation. This resulting difference is the studentized deleted residual (or simply the deleted residual).

The deleted residual for a given observation is typically larger than the regular residual statistic calculated using all observations, by construction, and is therefore referred to as an inflated predicted error. The quantile statistic is unit free, an advantage compared to the deleted residual, and the value corresponds to a p-value for the t-test of the hypothesis: H_o: the observation represents performance described by the mathematical model developed and applied to the remaining n-1 observations, versus H_a: the observation represents performance below the level of the mathematical model. Stated differently, the quantile statistic is the probability of using input $x_i \leq x$ with the level of all other inputs held constant, and producing the level of output predicted by the mathematical model considering only $n-1$ observations.

The calculation of the deleted residuals for a set of firms would seem to require processing $n+1$ regression if there were n units in the set. Although by today's standards regression estimations can be run rather quickly and hence it would not take much time to process $n+1$ regressions, it is shown in Neter, Wasserman and Kutner

[1985] the information in the full standard least squares regression can be used to calculate the ith deleted residual. The regression coefficients are computed by solving the normal equations

$$X'X\beta = X'Y \tag{6.17}$$

The projection matrix is defined as

$$H = X\left(X'X\right)^{-1}X' \tag{6.18}$$

and a diagonal element of this matrix can be found by

$$h_{ii} = X_i'\left(X'X\right)^{-1}X_i \tag{6.19}$$

If the fitted values are computed as

$$\hat{Y} = X\hat{\beta} = X\left(X'X\right)^{-1}X'Y = HY \tag{6.20}$$

then the standard residuals can be calculated as

$$e_i = Y - \hat{Y} \tag{6.21}$$

and the ith deleted residual is computed by

$$e_{(i)} = {e_i}\Big/{\left(1-h_{ii}\right)} \tag{6.22}$$

Typically the inputs and outputs describing the unit are assumed to be normally distributed and thus the n deleted residuals are distributed as multivariate normal. However, Cook and Weisberg [1982] have shown that the deleted residuals are typically effective even when the normality assumptions are violated mildly. The variance estimate of $e_{(i)}$ is:

$$s^2\left(e_{(i)}\right) = MSE_{(i)}\left(1+h_{ii}\right) \tag{6.23}$$

where $MSE_{(i)}$ is the mean square error when the i[th] observation is omitted in fitting the regression function, and $X_{(i)}$ is the X matrix with the i[th] observation deleted. Beckman and Trussell [1974] showed that the MSE for the full regression model can be simply expressed as a linear function of any jackknife variance estimate:

$$(n-k)\,\mathrm{MSE} = (n-k-1)\,MSE_{(i)} + e_i^2 \Big/ (1-h_{ii}) \tag{6.24}$$

Using this result, the standardized (or studentized) deleted residual is

$$t_{(i)} = \frac{\left(Y_i - e_{(i)}\right)}{s^2\left(e_{(i)}\right)} = e_i \sqrt{\frac{n-k-1}{SSE(1-h_{ii}) - e_i^2}};\ \text{for } i = 1,...,n \tag{6.25}$$

where k represents the number of terms in the linear regression model, and SSE is the regression sum of squared errors for the full regression model, which has $n-k$ degrees of freedom. From each reduced model, the standardized deleted residual has a t-distribution with $n-k-1$ degrees of freedom; thus the quantile statistic for the i[th] deleted residual is

$$\hat{p}_{(i)} = \Phi_{n-k-1}\left(t_{(i)}\right) \tag{6.26}$$

where Φ_i denotes the cumulative distribution function for the t-distribution with i degrees of freedom.

Quantile scores are relative scores, not absolute scores, so direct comparison is not possible. If unit 1 has a score of 0.66 and unit 2 has a score of 0.50 this does not imply that unit 1 is 16% better than unit 2. However, by constructing confidence intervals, comparison is possible. A covariance term can be calculated by

$$\mathrm{Covariance}\left(e_{(i)}, e_{(j)}\right) = \sigma^2 \frac{1-h_{i,j}}{\left(1-h_{i,i}\right)\left(1-h_{j,j}\right)} \tag{6.27}$$

where σ^2 is the variance of the error term. The interval can be calculated as

$$\left(e_{(i)} - e_{(j)}\right) \pm t_{1-(\alpha/2)}\left(n-k-1\right)\sqrt{s_i^2 - s_j^2} \tag{6.28}$$

where k is the number of terms in the linear regression model and $t_{1-(\alpha/2)}\left(n-k-1\right)$ is the $1-\left(\alpha/2\right)$ quantile of the t-distribution with $n-k-1$ degrees of freedom. Two observations are not significantly different if the confidence interval contains 0.

Efficiency estimates try to quantify the performance of a unit relative to a most productive state. In theory, the relationship between the location of the most productive frontier and the unit under evaluation could be measured in a variety of ways. The focus of quantile analysis is different; it is to quantify the relationship between a unit under consideration and its peers. There is no attempt to identify a most productive frontier or a maximum achievable output level based on a unit's input consumption using the quantile method. Thus typical assumptions made about the production process as outlined in Shephard [1970] or Simar and Wilson [2000] are not necessary. This is one reason quantile statistics can be said to depend less on modeling assumptions. Further because an efficient frontier does not need to be identified, less data is required by MQBA. Generally models requiring fewer assumptions are considered more robust.

To apply the quantile technique a regression equation and a functional form must be specified. The flexibility of the comparison is limited to some extent by the form selected; however, by selecting a flexible form, such as translog the trade-off between inputs and outputs and the return to scale properties do not need to be specified. In addition, by construction, quantile scores make full use of the range between 0 and 1. Other methods such as SFA enforce a skewed distribution or in DEA a certain number of units are needed to construct the frontier and thus often a significant proportion of units receive efficiency estimates of one. These benefits are traded off against the drawback of not identify the efficient frontier and thus absolute efficiency cannot be estimated.

6.6 Example

To demonstrate the MQBA technique the Charnes, Cooper and Rhodes [1981] (CCR 81) data set will be used. SFA and DEA calculation will also be shown for comparison purposes. The CCR 81 data set was constructed to evaluate a large-scale social experiment in public school education called Program Follow Through (PFT). The objective of CCR's analysis was to determine if PFT was improving the education achieved its participants. Before this could be determined the managerial efficiency of the schools needed to be controlled for and this is where DEA was applied. The outputs selected by CCR from a set of 11 possible outputs were

Y1: Total Reading score as measured by the Metropolitan Achievement Test

Y2: Total Mathematics score as measured by the Metropolitan Achievement Test

Y3: Coopersmith Self-Esteem Inventory, intended as a measure of self-esteem.

The inputs identified among a set of 25 possible inputs were

X1: Education level of mothers as measured in terms of percentage of high school graduates among female parents.

X2: Highest occupation of a family member according to a pre-arranged rating scale.

X3: Parental visit index representing the number of visits to the school site.

X4: Parent counseling index calculated from data on time spent with child on school-related topics such as reading together, etc.

X5: Number of teachers at a given site.

Data were collected for 70 schools, 49 of which participated in PFT and 21 of which did not participate in PFT. For each school, a sample of 100 students was taken and the

average is the data given for each input and output except X5. X5 is the total count of teachers working at a particular school.

For this data set the constructed variable method with the Cobb-Douglas functional form augmented by squared terms of output was selected for both SFA and the MQBA. As noted by Klein [1962] the curvature of the Cobb-Douglas function in multiple output space is incorrect because the elasticity of substitution between the outputs is constant. This has lead some to say the Cobb-Douglas transformation function is not an acceptable model for production with multiple outputs. Thus the squared terms of output were added. The resulting regression equation had nine terms

$$\ln\left(1/X_1\right) = x_2^* + x_3^* + x_4^* + x_5^* + y_1 + y_2 + y_3 + y_1^2 + y_2^2 + y_3^2 + \varepsilon \tag{6.29}$$

where lower case letters indicate the natural log of the variable and x_1 was the variable chosen as the dependent variable. The notation x_i^* represents the normalized value of the ith input relative to X_1. This equation was then used for both the SFA and revised QBA.

The results for four methods (SFA, DEA, DEA-W and MQBA) are shown in Table 6.1. At this point it is worth reiterating that MQBA estimates the proportion of the general population of firms that would require no less input than the ith firm, given the same set of output values produced. In contrast, the other methods attempt to measure the distance from a frontier representing the minimum input that can be achieved for a given level of output. Thus, the value of the efficiency scores would be expected to vary between MQBA and SFA or DEA. This is shown in the histogram of SFA and MQBA scores in figure 6.1.

Table 6.1. Results for Example from Charnes Cooper and Rhodes (1981)

Firm	SFA Efficiency	SFA Rank	DEA Efficiency	DEA Rank	DEA-W Efficiency	DEA-W Rank	Quantile Rank Score	Quantile Rank
49	0.941	13	1	1	0.9408	28	0.989	1
55	0.997	3	0.9994	28	0.9525	25	0.950	2
69	0.804	46	1	1	1	1	0.930	3
25	0.999	2	0.9787	32	0.8726	42	0.926	4
70	0.982	8	0.964	35	0.8335	52	0.922	5
22	0.987	5	1	1	0.9615	23	0.895	6
27	0.983	7	1	1	1	1	0.889	7
32	1.000	1	1	1	1	1	0.881	8
58	0.946	11	1	1	1	1	0.869	9
30	0.991	4	0.8934	59	0.7615	66	0.849	10
44	0.985	6	1	1	1	1	0.846	11
62	0.891	27	1	1	1	1	0.842	12
16	0.940	14	0.9501	43	0.8618	44	0.834	13
60	0.951	10	0.9804	31	0.8871	37	0.830	14
21	0.919	18	1	1	1	1	0.823	15
65	0.945	12	0.9754	33	0.9086	34	0.808	16
68	0.908	20	1	1	1	1	0.789	17
6	0.928	15	0.9099	55	0.8526	48	0.778	18
67	0.858	31	0.9462	45	0.8161	53	0.765	19
24	0.904	22	1	1	1	1	0.749	20
33	0.866	30	0.9521	42	0.8551	46	0.738	21
53	0.856	32	0.8696	63	0.793	57	0.730	22
41	0.901	24	0.9523	41	0.8398	50	0.714	23
52	0.904	23	1	1	1	1	0.673	24
20	0.905	21	1	1	1	1	0.664	25
28	0.964	9	0.9903	29	0.9362	29	0.655	26
18	0.804	47	1	1	0.9431	26	0.644	27
39	0.914	19	0.9415	47	0.9112	33	0.644	27
23	0.846	34	0.9748	34	0.8834	40	0.636	29
42	0.920	17	0.9531	39	0.9168	31	0.606	30
47	0.900	25	1	1	1	1	0.605	31
11	0.890	28	1	1	0.9791	22	0.604	32
12	0.855	33	1	1	1	1	0.602	33
66	0.808	45	0.9356	49	0.7563	67	0.590	34
3	0.842	36	0.9348	50	0.894	35	0.576	35
19	0.842	37	0.9526	40	0.8867	38	0.537	36
45	0.927	16	1	1	0.9587	24	0.528	37
51	0.751	61	0.9199	53	0.8056	55	0.502	38
4	0.834	38	0.9016	57	0.7636	65	0.457	39
2	0.816	44	0.901	58	0.8573	45	0.449	40
64	0.831	41	0.9303	51	0.7861	59	0.433	41
8	0.798	48	0.905	56	0.8014	56	0.422	42
31	0.822	43	0.8369	69	0.7169	69	0.400	43
26	0.796	49	0.9425	46	0.8692	43	0.383	44
10	0.789	50	0.9408	48	0.8906	36	0.362	45
34	0.779	56	0.859	66	0.7896	58	0.351	46
7	0.833	40	0.8914	61	0.7477	68	0.340	47
57	0.781	54	0.9269	52	0.8544	47	0.339	48
17	0.831	42	1	1	1	1	0.338	49
63	0.843	35	0.9634	36	0.8759	41	0.323	50
37	0.833	39	0.8393	68	0.7802	63	0.315	51
9	0.786	52	0.8585	67	0.7691	64	0.304	52
43	0.787	51	0.8647	64	0.7843	62	0.277	53
54	0.763	58	1	1	1	1	0.268	54
40	0.761	59	0.9498	44	0.8838	39	0.258	55
13	0.760	60	0.8623	65	0.7845	61	0.256	56
29	0.880	29	0.8833	62	0.8368	51	0.221	57
61	0.894	26	0.8927	60	0.8062	54	0.169	58
1	0.721	63	0.9621	37	0.9214	30	0.122	59
46	0.676	66	0.9129	54	0.7855	60	0.108	60
14	0.774	57	0.9897	30	0.9136	32	0.079	61
59	0.783	53	1	1	1	1	0.064	62
38	0.747	62	1	1	1	1	0.053	63
35	0.689	65	1	1	0.9424	27	0.044	64
56	0.562	68	1	1	1	1	0.042	65
50	0.640	67	0.9587	38	0.8483	49	0.039	66
48	0.417	70	1	1	1	1	0.017	67
36	0.693	64	0.7929	70	0.6898	70	0.013	68
5	0.781	55	1	1	0.9856	21	0.008	69
15	0.481	69	1	1	1	1	0.003	70

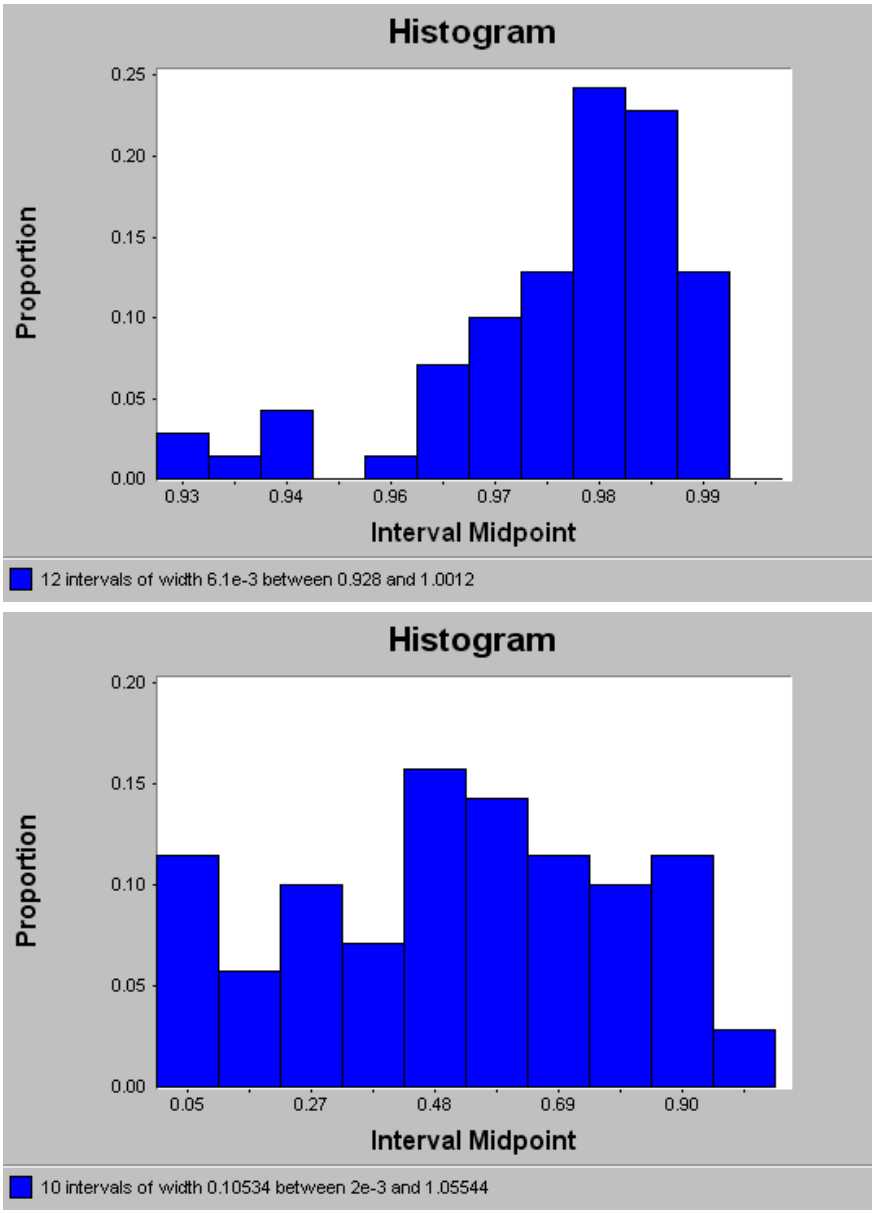

Figure 6.1. The top graph show the distribution of SFA estimates while the lower graph shows the distribution of MQBA efficiency estimates.

In table 6.2 the Spearman correlation coefficient between each of the three methods is shown. While the SFA and QBA efficiency scores are highly correlated, they both vary

markedly from the DEA efficiency estimates. This is a typical result observed in many papers such as Bauer et al. [1998] among others.

Table 6.2. Spearman's Correlation Coefficients

	QBA	SFA	DEA
SFA	0.864	1.000	0.252
DEA	0.224	0.252	1.000
DEA-W	0.182	0.176	0.933

One intuitive reason DEA results vary from SFA or MQBA is the weight selection issue in DEA. The DEA method searches over all possible weightings of inputs and outputs to find the set of weights that allows any individual firm to achieve the highest possible efficiency score. This process allows a firm to select its best input and best output or any combination of inputs and outputs to determine its efficiency estimate. Thus while performing poorly on any criteria involving certain inputs, in DEA those inputs for which a firm did not perform well can be given smaller weight and the input and output metrics for which performance is good can be given larger weight.

However, this is not the case for SFA and revised QBA. When a regression equation is specified, coefficients of the independent variables are determined and the same values for these coefficients are used in the evaluation of all units. Thus, a firm's worst characteristics cannot be ignored. For comparison purposes DEA-W, a DEA model with weight restrictions forcing each output and input to contribute at least 10% to the product of weights times input or weights times outputs, is also shown.

Notice that in DEA-W, as each input and output is forced to contribute to the efficiency calculation, the correlation between DEA-W and MQBA or SFA is lower than

the correlation of DEA to these values. This contradicts the previous argument. Upon further investigation we find monotonicity is violated for outputs 1 and 3 in the regression estimation. Thus the restricted DEA frontier is further way from the SFA frontier than the unrestricted DEA frontier. An example of this can be shown in figure 6.2.

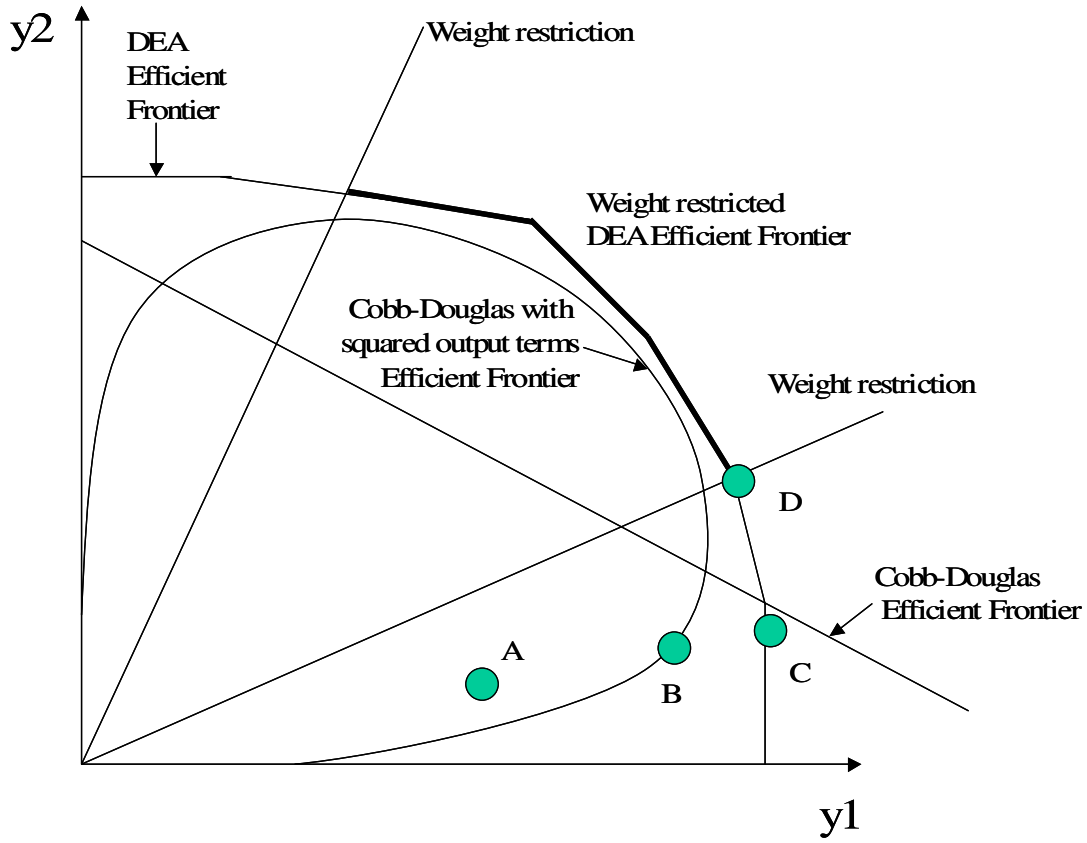

Figure 6.2. The DEA efficient frontier drawn in two dimensions with weight restrictions and the Cobb-Douglas efficient frontier imposed over top.

It appears that weight selection alone cannot explain the differences between DEA and SFA or QBA.

An alternative explanation is the presence of outlying observations. While SFA and QBA are fairly robust to outlier data using the residual term or a component of the residual term to model error in measurement, DEA does not have this attribute. An outlying observation can impact the distribution of efficiency estimates, not only by having an inaccurate efficiency value, but also by changing the shape of the efficient frontier and impacting other efficiency estimates, by downwardly biasing them.

In a recent paper discussing outlier detection, Johnson and McGinnis [2005a] demonstrate how to identify outliers. When the iterative multi-orientation method (requiring an outlier to be an outlier from both an input and output perspective) is applied to this data set, 4 outliers are identified. If the DEA, SFA, and MQBA estimates are recalculated for the smaller data set, the correlations are shown in table 3.

Table 6.3. Spearman's Correlation Coefficient

	QBA	SFA	DEA
SFA	0.932	1.000	0.317
DEA	0.228	0.317	1.000
DEA-W	0.193	0.253	0.913

The correlation between SFA and DEA or DEA-W has increased, but the correlation between DEA or DEA-W and QBA is relatively unchanged.

A further explanation of this difference can be attributed to additional flexibility in the parametric models. While it is typically argued that DEA is more flexible because it does not require the assumption of a functional form, it does require all variables be identified as either inputs or outputs prior to the analysis. The assumption is that each input can substitute for other inputs and each output can substitute for other outputs. This

is not true for the MQBA method. If the input oriented distance function is used, as shown in the example, an arbitrary input is selected to be the dependent variable in the MQBA method. The other inputs which were not chosen as the dependent variable can have a positive or negative coefficient. A positive coefficient indicates the inputs are compliments, which is not possible in DEA.

In the multiple output / multiple input SFA a single input or output is selected as the dependent variable. The error term and thus efficiency is measure solely in the dimension of the dependent variable. In DEA, for an input oriented analysis, the efficiency is measured as a distance in all input dimensions. This difference in dimensionality causes further diversion between DEA efficiency estimates and regression based methods.

Seventy observations were in the CCR81 data set; however, this is just a sample of the schools that were candidates for the PFT. Without collecting the data on the other schools it is impossible to know if this sample correctly constructs the efficient frontier for all candidates for the PFT. In the sample of 70, if there were more observations randomly selected from a certain region of the input/output space and fewer observations taken from another, this may cause a bias in both the SFA and DEA efficiency estimates. The region with more observations may under-estimate efficiency relative to an area where a proportional number of observations to the number of schools among all schools was taken. Similarly, a region with fewer observations may over-estimate SFA and DEA efficiency scores. This problem arises when an efficient frontier is estimated and the bias cannot be observed without collecting data on the entire population. Thus, while MQBA cannot offer absolute efficiency estimates, the relative efficiency estimates would not be

susceptible to this problem. MQBA efficiency estimates are different from the values estimated in SFA and DEA in this sense.

Further, the MQBA estimates are bounded by zero and one and range over the entire interval by construction, the proportion of observation in any interval on the range is determined by the data; whereas, in SFA this is not true because the distribution of efficiency scores is defined in the model. In DEA an observation that has the highest ratio of output to input for any pair of inputs and outputs is scored as efficient. This often causes a significant proportion of observation to be scored as efficient. Recent research results show the distribution of efficiency estimates is impacted in a systematic way by the dimensionality of the model Johnson and McGinnis [2005b]. Thus the MQBA efficiency scores cannot give any insight into absolute efficiency levels because an efficient frontier is not constructed, but MQBA's characterization of the distribution of efficiency estimates are less affected by modeling assumptions relative to SFA or DEA.

6.7 Conclusion

This paper has shown how the quantile-based analysis can be expanded to simultaneously handle multiple inputs and multiple outputs without using an arbitrary weighted aggregation technique. The analysis has several advantages and restrictions relative to other performance evaluation analysis such as DEA or SFA. It allows the ranking of firms according to efficiency and mitigates the effects of outliers. Further, the method is relatively easy to implement.

This method is particularly useful for cross-sectional data. There is extensive research to reduce the assumptions necessary in stochastic frontier analysis (such as those described in Kneip and Simar [1996]); however, many of them require panel data.

Therefore, the simultaneous multi-input multi-output quantile based analysis method is suggested as an alternative to DEA or SFA in cross sectional data and when the estimation of a frontier is not possible do to limited data.

CHAPTER 7

PRODUCTIVITY MEASUREMENT IN THE WAREHOUSING INDUSTRY

7.1 Introduction

In this chapter, data on warehouse performance collected over the past 5 years will be used to benchmark the performance of each observation against the other observations in the data set. The performance measurement technique of data envelopment analysis (DEA) will be used. While using DEA, several benefits and shortcomings have been identified. Methods developed in the previous chapters for improving the application of DEA are applied in analyzing the warehouse data. It will be demonstrated that the results of an efficiency study of warehouses are significantly different by using the techniques suggested in the previous chapters.

A goal of our analysis of this data set is to move beyond the single factor productivity measures which are pervasive in the warehousing industry. Perhaps the most common is lines shipped per labor hour. This type of measure is not very informative because of the possibility for input substitution. For the same output levels, warehouses with more capital may be able to use less labor because the capital substitutes for labor. Therefore, it is desirable to use a measure that considers the multi-input / multi-output nature of warehousing.

In warehouses, the prices of outputs, and occasionally inputs, cannot be measured accurately excluding the use of economic efficiency measurement. These outputs, such as lines shipped or orders shipped often are not sold; thus, the value of these outputs is

not clear. Although accounting data might be used when measuring the value of inputs, accounting data can provide a poor approximation for economic prices (i.e. marginal opportunity costs), because of debatable valuation and depreciation schemes. When DEA is used in the most basic production model, it does not require price or cost information, which matches our data availability.

DEA is a non-parametric efficiency estimation method based on minimal prior assumptions about the production possibility set. This is an important characteristic because economic theory does not put forth strong hypotheses about the production function, and in many cases reliable production function specification tests are not available. The boundary of the production possibility set is called the efficient frontier and characterizes how the most efficient firms tradeoff inputs and outputs. Further DEA allows each production unit to determine the value of each input and output individually. In contrast, regression based techniques for approximating the production function would estimate values of inputs and outputs based on the entire data sample, and thus would require identical marginal rates of substitution between inputs and outputs for all production units.

DEA efficiency measures tend to over-estimate efficiency and would only under-estimate efficiency in the cases of data entry error or misidentification of the peer group. The minimal convex hull which encompasses the data and maintains the assumptions about the production possibility set is used to estimate efficiency. Thus, efficiency estimates are optimistic in the sense that a unit can only have a worse efficiency estimate with the inclusion of additional data.

Classic microeconomic theory assumes all units are efficient; otherwise, they would go out of business. Therefore, models based on this theory do not allow for inefficiency and would not provide insight into the problem of measuring performance. While modern economic theory has put forth some theories which allow for inefficiency, such as principle-agent theory, these theories are not strong enough to justify a particular statistical distribution for efficiency. It is the lack of information about the distribution of efficiency coupled with assumption of constant marginal rates of substitution across all facilities which lead to our selection of DEA over its most noteworthy competitor stochastic frontier analysis (SFA).

7.2 The Process of Applying Data Envelopment Analysis in Performance Analysis

While using DEA to investigate the performance of warehouses, we discovered areas in which DEA fails to model our problem sufficiently. In these cases we document the progress we have made to develop methods in order to overcome the issues and note areas in the literature in which others have attempted to address these problems.

The most fundamental problem is to understand what we are quantifying when measuring deviations from the frontier. There are three common explanations in the literature related to this problem: model misspecification, data problems, and truly non-optimizing behavior.

Model misspecification is perhaps the most accepted. It is not always clear what inputs or outputs should be used and which should be left out. Sometimes it is not clear whether a particular variable is an input or an output Cook and Zhu [2005]. Obviously, if

the model is incorrectly specified, inefficiency may be indicated by the analysis even if all units are behaving efficiently.

Errors in data can lead to incorrect estimates of efficiency and possibly indicate inefficiency in cases where it does not exist. DEA assumes all production units are measured perfectly, and the definition of each input and output is understood and measured in the same way for each unit.

Finally, it is possible that a firm is, in fact, behaving in a non-optimal manner. This would contradict the classical microeconomic theory. However, there are reasons to believe this could be possible. First, while the theory requires inefficient firms to go out of business, it does not say how quickly. Therefore inefficient observation may be DMUs that will eventually go out of business, but have not yet. It is also possible an agency problem exists when the goals of the production unit, as a whole, are not the same as the goals of an individual manager or employee. Therefore, a production unit may be observed to be acting inefficiently based on the separation of the level being observed (production unit level) and the decision level (employee level). Further, long-term decisions are made, such as facility purchases, facility layout, or equipment purchases that lock a production unit into a limited course of action in the short-run. Thus, as future events unfold, it is possible that inefficient performance results because forecasted events did not occur. When measuring production efficiency, it is often hard to separate the effects of accuracy in forecasting and production performance.

In our application, we recognize that each of the explanations, to some extent, is valid. Our model is incomplete, but it is designed to capture the most important inputs used and outputs generated by a warehouse. Because there may be errors in the data we

have collected, we have developed an outlier detection technique which will be summarized in the following section. Finally, we believe there is non-optimizing behavior in warehouses.

This behavior may be a result of the agency problem or warehousing being a dynamic environment where some warehouses are able to implement new operational methods or technologies before other warehouses. This may be due to knowledge advantages that one DMU might have over another, or it might be related to the latency of decisions such as a given DMU may have made a prior decision and invested a sum of money. The DMU is not willing to change the production method or install new equipment without first attempting to receive some value from the previous decision. Because of the latency issue, it is not possible to separate forecasting ability from production efficiency in cross-sectional data. Thus we proceed to measure efficiency with this understanding of efficiency in warehouses.

Beyond the definition of efficiency, DEA is based on a set of assumptions which are considered not to be intrusive. There are at least four assumptions when adopting the DEA paradigm: (1) the proper orientation for measuring efficiency can be selected, (2) the assumptions about the production possibility set hold, (3) the observations give a good representation of the complete production technology, and (4) the observations are measured accurately. Each of these assumptions will be further described and the reasons why these assumptions are appropriate in this analysis will be discussed.

7.2.1 Orientation for Measuring Efficiency

Orientation refers to the direction taken for measurement from an observation under evaluation to the efficient frontier of the production possibility set. There are four common orientations: directional, input, output, and hyperbolic.

Directional orientation, developed by Chambers, Chung and Fare [1996], simply adjusts inputs and outputs to move the DMU to the efficient frontier. The directional orientation is the most general and leaves the user with complete flexibility to define the direction in which to measure. However, this added flexibility is hard to utilize because there is rarely any information to help the analyst make such a decision in anything but an ad-hoc manner. For example, a direction chosen may model the case where some input levels or output levels can be adjusted, while other input and output levels cannot. Estimating the degree to which an input or output level is flexible can be difficult. Thus, even though the other orientations are special cases, they are still frequently used because there is a common reasoning behind their selection.

The input orientation is an equaproportional contraction of all inputs, holding outputs constant, which moves the DMU to the efficient frontier. Therefore, if a production unit could argue that they have complete control over the acquisition of inputs, but are not able to influence output levels (demand is exogenous), then an input orientation may be justified.

Similarly, an output orientation is an equaproportional expansion of all outputs, holding inputs constant, which moves the DMU to the efficient frontier. A production unit that completely controls its output production but cannot change input levels, due to

stickiness in the input markets or corporate control of input acquisitions, can select an output orientation to model these limitations.

Finally, the hyperbolic orientation is a simultaneous equaproportional expansion of outputs and contraction of inputs. This allows a production unit to consider adjusting both input and output levels and determine the level that will allow them to become efficient with the smallest proportional change of all variables. However, this orientation assumes all inputs and outputs are adjustable, which may not be true in all industries or for certain models.

The orientation selection is important for the accuracy of the efficiency estimates. Although it has not been explored in the literature, it is also possible to imagine various orientations may be appropriate for production units within the same industry using the same model due to reasons such as labor conditions (the local market for labor or unionization) or firm structure (distributed decision-making about production unit investment decisions). It may be easier for one firm to adjust certain inputs or outputs than another firm in the same industry.

For our warehousing data set, we want to measure the efficiency of warehouse operations from the warehouse manager's perspective and determine the orientation based on the following reasoning. Warehouses are often considered to be cost centers which typically do not generate profit, but rather aid in the distribution of goods, preferably at minimal cost. It is very difficult to estimate the value of the outputs generated by the warehouses. The warehouse manager makes decisions about the warehouse's operations, such as labor assignment, operation methods, and equipment needs. However, the inbound and outbound flows of goods are typically beyond the

warehouse manager's control. These flows are affected by customer demand, advertising, pricing, and procurement. An input orientation is taken for this analysis because the required outputs of the warehouse are defined externally, and it is the warehouse manager's responsibility to fulfill these requirements using minimal resources.

7.2.2 Assumptions about the Production Possibility Set

The construction of the production possibility set is based on several assumptions. Most are considered rather weak and have been accepted by many analysts for approximately 50 years. However, recently the validity of these assumptions has become a topic in the literature. Banker, Charnes and Cooper [1984] summarizes these maintained assumptions as:

1. Data envelopment (DE): If a set of n comparable firms, call this $S \equiv \left\{\left(x_j, y_j\right)\right\}_{j=1}^{n}$, is defined and the production possibility set is called T, then data envelopment implies $S \subseteq T$.
2. Graph convexity (GC): $T = co\left(T\right)$, with $co\left(T\right) \equiv \left\{\left(\lambda x + \left(1-\lambda\right)x', \lambda y + \left(1-\lambda\right)y'\right) : \left(x, y\right), \left(x', y'\right) \in T, \lambda \in \left[0,1\right]\right\}$ for the convex hull.
3. Ray unboudedness (RU): $T = c\left(T\right)$ with $c\left(T\right) \equiv \left\{\left(kx, ky\right) : \left(x, y\right) \in T, k > 0\right\}$ for the conical hull.
4. Strong disposability (SD): $T = m\left(T\right)$, with $m\left(T\right) \equiv T + \mathbb{R}_{+}^{M} \times \mathbb{R}_{+}^{S}$ for the monotone hull.

The authors then define the minimum extrapolation principle to the maintained assumptions and obtain the production possibility set as the conical convex monotone hull of the observations:

$$c\left(co\left(m\left(S\right)\right)\right) \equiv \left\{\left(x, y\right) : x \leq X\lambda, y \leq Y\lambda, \sum \lambda \in R_{+}^{n}\right\}$$

As noticed by Afriat [1972] and Banker, Charnes and Cooper [1984], if the ray unboundedness assumption is relaxed, while maintaining the graph convexity, this allows the production possibility set to model variable returns to scale. However, Petersen [1990] noted that GC enforces marginal products (or the increase in output per unit of input) to be non-increasing. The typical explanation for increasing marginal products is the concept of specialization. By excluding this phenomenon, the production possibility set is not consistent with typical understanding of variable returns to scale.

Cherchye and Post [2003] note that convexity assumes the following do not exist: (1) indivisible inputs and outputs, (2) economies of scale, and (3) diseconomies of scope. A quote from Farrell [1959] elaborates this point,

> 'A glance at the world about us should be enough to convince us that most commodities are to some extent indivisible and that many have large indivisibilities. Similarly, whenever one refers to "economies of scale" or of "specialization", one is pointing to concavities in the production functions. There is thus no need to argue that importance of either indivisibilities or concavities in production functions – the former are an obvious feature of the real world, and the latter have constituted a central topic in economics since the time of Adam Smith.'

For these reasons the assumption of convexity of the production possibility set may warrant further investigation, and alternative models relaxing this assumption would be valuable in these studies.

For warehousing, convexity may be a reasonable assumption. Additions of an individual unit of equipment, personnel or space do not often result in significant increases in output. Thus indivisibility of input is rarely significant. Further the most common outputs generated by a warehouse are lines shipped, orders, and accumulation. All of these typically are measured in thousands or tens of thousands; therefore the indivisibility of a single output unit is relatively small to the overall value. Based on these observations, indivisibility does not seem to be a significant issue in warehousing.

Specialization implies that increasing returns is necessarily associated with a change of method. But this implies there are indivisibilities in production. In other words, the specialized tasks available at large scale are not available at the smaller scale; consequently, as the scale of production increases, these indivisibilities are overcome and methods not previously available become available. However, indivisibilities have been argued to be insignificant in this warehousing analysis. Thus specialization will be assumed to be of minimal concern; in turn convexity of the production possibility set will be assumed.

Strong disposability eliminates the possibility of extreme congestion, i.e., the situation in which marginal productivity rates are negative. By assuming strong disposability, extreme congestion is not considered. This has been discussed extensively by Fare and Svensson [1980], and an alternative model without strong disposability has been developed by Fare and Grosskopf [1983]. However, if a warehouse is expanding or considering expansion, before extreme congestion can occur, the warehouse will go through a congested period, i.e., marginal productivity will decline. It is assumed for this study that no warehouses are operating under extreme congestion.

The assumption of data envelopment has not been challenged except to say that each data point may not have been measured accurately. Observations with errors in measurement should be treated differently. This topic will be discussed in section 7.4.

7.2.3 The Set of Observations Gives a Good Representation of the Complete Production Technology

When using DEA a concern is, "do the observations give a good representation of the complete production technology, or in other words, how many data points, if data is drawn randomly, are required to accurately describe the production frontier in DEA?" If the answer is "yes," a user of DEA will have a statistical confidence interval for the efficiency estimates calculated based on the amount of data used in the calculation. We can expect the complexity of the model to complicate this issue.

As the size of the model increases, the dimensionality of the production frontier will also increase. When using DEA, the production frontier is identified as a set of efficient units and their convex combinations. Therefore, as the dimensionality of the production frontier increases, typically the amount of data needed would also increase. However, this is not a strict rule. It is always possible to have one very efficient unit in all dimensions and only that point would be necessary to define the production frontier. For a review of this literature refer to section 3.8.

The data required to define the production frontier can vary widely depending on the relative structure of the units being compared. For example, in an industry in which there is an obvious leader who is the best in all respects, a single data point may suffice to define the frontier. However, if the industry is very competitive and is competing in many different measures of output, we may see many more efficient units. Thus we must

be aware, during any analysis using DEA, the method is limited by the group of units being analyzed. It is not possible to use only one data set and make conclusions about DEA. A structured analysis of the data requirements for DEA is lacking in the literature.

Using the outlier detection method described below, a preliminary investigation into the quality of the production technology representation can be undertaken. The method considers overly efficient and overly inefficient observations. If the data are not well clustered, even when the dimensionality of the model is reduced, a large proportion of the data points will be flagged for further investigation. If a significant proportion of observations are flagged as outliers, the model should be reconsidered, or the data may not be sufficient to represent the production technology. In this case stronger assumptions can be made about the production technology to allow investigation into smaller data sets.

Our findings are that the data requirements are typically much higher than the rules of thumb suggested in the literature, even for very small models. We also have developed an analysis method called the multi-input / multi-output quantile-based method (MQBA). This method allows an efficiency ordering to be estimated in smaller data sets by assuming a flexible form of the production function and uses a distance function to aggregate either the inputs or the outputs so that a multi-input / multi-output model can be estimated.

7.2.4 The Observations are Measured Accurately

To better understand the quality of a group under analysis and to partly address the issue of identifying observations that may not have been measured accurately, we developed an outlier detection method. Here we will adapt the definition of outlier

provided by Gunst and Mason [1980], "as observations that do not fit in with the pattern of the remaining data points and are not at all typical of the rest of the data". Thus once an outlier is detected, it must be further analyzed to determine the reason it was identified. A decision needs to be made to include or exclude the observation from further analysis.

Barnett and Lewis [1995] suggests four reasons outliers exist: deterministic, incorrect expectation, slippage and contamination. Deterministic would be simply an incorrect measurement. These types of outliers should be excluded from further analysis, or the data should be corrected. Incorrect expectation would be an improper assumption about the range or the variance of the variable under consideration. These expectations are often a part of outlier detection methods. Thus when an outlier is found, the analyst should consider if this data truly contains invalid information or if the detection method includes an improper assumption. Slippage implies over time the range of a variable may change. The variables measured initially would not be comparable to the variables measured at the end of the experiment. In a production context this might imply technical progress meaning a production unit can be more productive because of technological improvements or learning by doing.

Finally, collections of observations under analysis may not truly be comparable. This is called contamination, when two separate groups are included in a single analysis. An example of contamination might be the analysis of efficiency of trucking companies from both the U.S. and Mexico. The differences in the quality of roads actually make these two types of trucking companies not comparable and would indicated that two distinct data sets are mixed, or one is contaminated by the other.

The outlier detection method described below can identify outliers related to deterministic, slippage, and improper expectations issues; however, it cannot detect the contamination problem. For the contamination problem, clustering analysis methods would be suggested such as those described in Hand [1997] or Hand [1981].

A previously developed outlier detection model, Wilson [1995], identifies outlier by their distance from the efficient frontier when the given observation is excluded from the reference set. Then these observations are prioritized by their impact on the efficiency scores of the remaining observations. While this method identifies outlier for having overly good performance, it does not identify errors for the excessively inefficient firms.

To address this issue, Johnson and McGinnis [2005a] developed an outlier detection method that uses super efficiency scores relative to an efficient frontier but also for an inefficient frontier and searches for outliers relative to both frontiers. It has been noted previously that the super efficiency can always be calculated by using the hyperbolic super efficiency measure, given all input data are positive numbers. As in any statistical method, Johnson and McGinnis's method requires the specification of a critical value (super efficiency level) beyond which observations will be flagged. If a large percentage of observations are flagged with relatively weak super efficiency levels specified, this may be an indication that the model of the inputs and outputs is poorly specified or the group of observations does not use very similar production functions. This will be illustrated on two different warehousing data sets, one large and one small.

7.3 Warehousing Data Analysis

The large data set consists of nearly 400 different warehouses collected over a five-year period, and will be treated as a cross section because the technical progress during

the five-year period is believed to be minimal. Much of this information has been collected via the iDEAs website, (www.isye.gatech.edu/ideas). This is a website that allows warehouse managers to log-in, enter data about their warehouse and receive a DEA variable returns to scale efficiency estimate based on set of warehouse data collected previously. The data have been collected with guidance related to the definitions of the values requested and contact information available to any user who wishes to have further assistance in entering their data. The initial model uses 5 outputs and 3 inputs. Warehouses actually use other input and outputs, but this set of 8 measures is believed to capture the most important inputs used and outputs generated. This model was initially developed by Hackman et al. [2001] and demonstrated on a smaller data set.

Even within warehouses there is still a great amount of variation in the type of product and the frequency of demand. Some warehouses are mostly Maintenance, Repair and Operating Supplies (MRO) and some are on-line retailers. The size of the products ranges from automobile spare parts to compact discs. The nature of order picking can vary dramatically.

We define a warehouses with a predominant pick mode to be warehouse that picks more than 80% of customer lines in a particular holding size, either broken case, full case or pallet. If a warehouse does not have a predominant picking mode, the warehouse is called a mixed picking warehouse. Within the 390 warehouses in this data set, 38% are predominately broken case picking, 15% are predominately full case picking, 7% are predominately pallet picking and the remainder are mixed picking. These various picking modes drive the operations of the warehouse.

The size (amount of physical volume) of a typical customer order impacts the storage, the picking and the labor utilization of a warehouse. The size of a customer order is highly correlated with picking mode. Even though all observations in the database are warehouses, various observations in this data set could be using different production functions and pick mode represents a way to identify separate groups. However, initially all warehouses will be placed in a single group and by applying the outlier detection method we attempt to identify inaccurate data that have been entered via the website and to understand the quality of the group of warehouses.

7.3.1 Outlier Detection Results

First, the critical level for the outlier detection method was investigated. The critical level is the percentage increase in inputs or percentage decrease in outputs necessary to move an observation under evaluation into the production possibility set constructed using DEA. A very loose critical level is 1.5 which corresponds to a 50% increase in inputs. If this observation, after increasing its inputs, is then located within the production possibility set, it is not flagged as an outlier. An observation that is still outside of the production possibility set after the increase in inputs is flagged. Stricter criteria can also be chosen. However, because of the method of collection, it may be expected that more than the rule of thumb, suggested by Barnett and Lewis [1995] of 10% of the population, could be outliers without indicating misspecification of the model or poor group identification. To begin, critical level values of 1.2 and 1.5 have been investigated. The results are shown in table 7.1.

Table 7.1: The impact of model size and critical value on the number of outliers

			Inputs x Outputs			
			3x1	3x3	3x4	3x5
Critical Super Efficiency Value	1.2	Count Kept	267	19	19	6
		% Kept	0.68	0.05	0.05	0.02
	1.5	Count Kept	343	216	125	68
		% Kept	0.88	0.55	0.32	0.17

Here the 1 output model uses lines shipped as the output, the 3 output model uses the types of lines shipped (broken case, full case, and pallet), the 4 output model excludes storage function from the full model, and the 5 output model include storage function. The outputs used in the various models indicate the relative importance of the outputs in explaining warehouse performance as the model size becomes smaller. Lines shipped are typically the best single output metric to use in assessing efficiency of a warehouse. Therefore, it is included in all models.

Other orderings of the outputs have been investigated with minimal change in the rate of decline of the number of observations retained as the model size grows. While not all outputs are produced by all warehouses, all inputs are used by all warehouses. Thus, for the investigation of the critical level, the number of inputs has been held constant. However, the model size effect on the number of observations kept in the analysis was further investigated by varying both the number of inputs and outputs. The results are summarized in tables 7.2 and 7.3.

When a multi-input / multi-output model is defined, 2 inputs and 3 outputs, 1/3 of the observations are flagged as possible outliers. This begins to characterize the amount of

Tables 7.2 and 7.3: The count and the percentage of the number of observations kept in the reference set for a variety of model sizes .

Count of Observations Kept

		Outputs			
		1	3	4	5
Inputs	1	354	305	228	189
	2	344	263	152	98
	3	343	216	125	68

Percentage of Observations Kept

		Outputs			
		1	3	4	5
Inputs	1	91	78	58	48
	2	88	67	39	25
	3	88	55	32	17

data necessary to approximate the efficient frontier in multi input / multi output models. The percentage of the observations kept drops rapidly for models larger than 2 by 3.

This analysis demonstrates there is a significant impact of model size on the number of outliers identified. For the multi-input / multi-output models the percentage of outliers identified is much larger the rule of thumb of 10% suggested by Barnett and Lewis [1995]. The literature on DEA also has two rules of thumb about the data requirements for a DEA analysis; however, these do not consider the presence of outliers. For a 3 by 5 model the two rules of thumb suggest 16 and 24 observations, both of which are less than

the number of observations remaining after the outlier detection process. Based on these suggestions the iDEAs data set could support a DEA analysis after outliers have been removed. It is also interesting to consider if a smaller group of warehouses handling mostly the same type of product would similar percentages of observations be flagged as possible outliers. This will be investigated in 7.3.2.

The distribution of efficiency scores was investigated. For any size model, the distribution of efficiency estimates has a large percentage of observations found to be efficient, typically on the order of 15 to 35% of the set under evaluation. The remaining observations were distributed with a mean in the range of 0.50 to 0.75 and a standard deviation in the range of 0.2 to 0.3 (see figure 7.1 and 7.2 for example distributions).

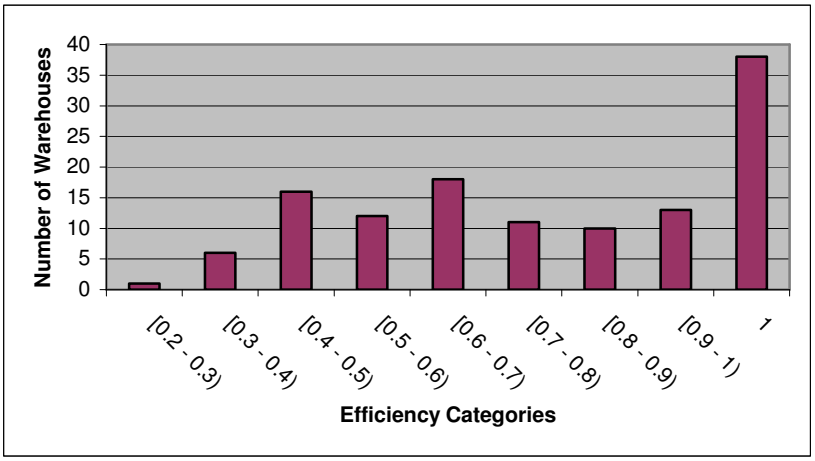

Figure 7.1. Histogram of Efficiency Estimates for a 3x3 Model

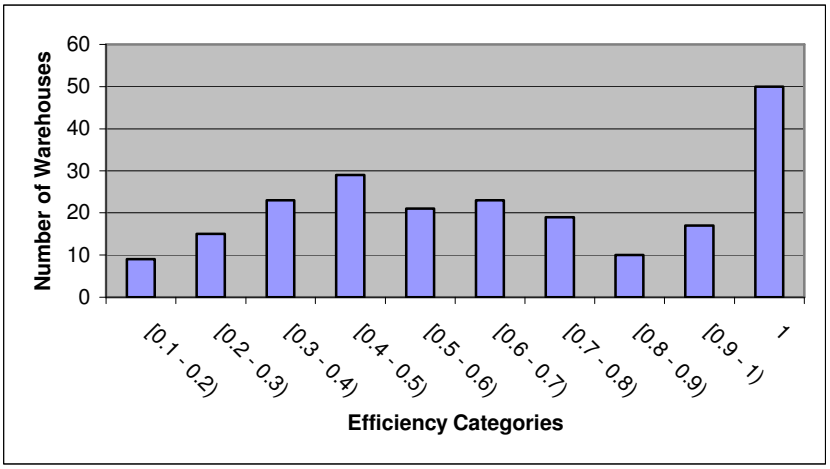

Figure 7.2. Histogram of Efficiency Estimates for a 3x4 Model

This result indicates there are not many observations that are inefficient but are close to the frontier (there are very few observations on the range of 0.7 to 0.9). This conclusion is similar to the results of Kittelsen [1999].

7.3.2 Industry Specific Study

In order to verify the results for iDEAs' analysis, a second set of data for a specific industry was collected and investigated. A trade association related to a particular industry was contacted to aid in the industry specific study. Data were collected for 25 warehouses in this industry. For this peer group, the units being handled in the warehouses should be similar. Presumably a more even mix of warehouses performing at all levels of efficiency is in the data set because there is no self-selection. The trade group was already formed; thus the problem of a warehouse performing at a particular level self selecting into the data set is avoided.

By applying the outlier detection method to this data set we found that 25 warehouses are not enough data for a model with multi-inputs and multi-outputs. For a summary of the outlier results see table 7.4.

Table 7.4: Observations Remaining after Flagged Observations are removed

		Output			
		1	3	4	5
Input	1	20	3	3	0
	2	17	0	0	0
	3	16	0	0	0

Perhaps using a 3x1 model would have characterized the substitution effects among inputs However, the differences in the amount of effort involved in shipping lines of the various holding sizes is still a concern. Thus a multi-input / multi-output model is desired. By using the multi-input / multi-output quantile based approach (MQBA) additional assumptions have been made about the functional form of the production function, which allows efficiency estimation for each observation. Using the quantile based approach, an ordering is created, and statistical significance of the difference between two estimates can be determined using a t-test.

7.3.3 Quantifying the Differences Between Estimated Performance With and Without Suggested Improvements

The methods for improving DEA have been described in the previous three chapters and the arguments related to production theory have been presented. The purpose of this section is to show that by applying these methods, results significantly different from standard DEA will be reached. This will be shown for the two data sets previously discussed. To the extent that these data sets represent typical data sets, this result can be

generalized. It is still necessary for the analyst to understand the production theory arguments and decide if the technique is appropriate for their analysis.

The outlier detection method described previously flags observation for further investigation. If the observation cannot be verified, it is removed from the analysis. This causes the data sets being compared, for a standard DEA model and the model augmented with the additional methods describe in this thesis, to be of different sizes. Because overly efficient outliers are present in the original dataset, the efficiency estimates of other observations are downwardly biased. Therefore an investigation of the two methods based on comparing differences in estimated efficiency may give misleading results.

A method is needed for comparing the results of the two analyses—conventional application of DEA versus the analysis using the methods developed in this thesis, referred to as the "augmented DEA," or aDEA. The following approach was taken. First, efficiency estimates were calculated using both methods. Then, observations identified as outliers in aDEA were removed from the conventional DEA results. Then, for the two sets, efficiency quartiles for each observation were calculated.

For the iDEAs data set, 53% of the observations had different quartiles for aDEA and standard DEA. For 17% of the observations the quartiles differed by at least 2 levels. For the industry specific data set, 64% of the observations changed quartiles, and for 28% of the observations, the quartiles changed by at least 2 levels.

These results show that the ordering of efficiency estimates is significantly different for the two methods. Because the augmented DEA methods are based on concepts from

production economics and are argued on a theoretical basis, the significant impact on efficiency results is strong argument for applying these methods.

7.4 Conclusions

This chapter has explored applying DEA and other productivity measurement techniques to warehousing data. A number of concerns related to the DEA model have been identified. In an attempt to address some of these concerns an outlier detection method has been applied. This method also gives information about the relative quality of the model, quality of the data, and the group identified for analysis. Based on these results it is clear the data requirements in DEA are much larger than the rules of thumb available in the literature. In our example of a 2 input / 3 output model, a data set of 390 observations is reduced to 263 based on the outlier detection technique. This implies that for one of the three reasons above (among quality of the model, quality of the data, and the group identified for analysis), nearly 1/3 of the observations are excluded from the analysis. This suggests that DEA analysis should not be used on small data sets.

As demonstrated with the industry specific data, a set of 25 observations taken from a well-defined group does not support a multi-input / multi-output model. The amount of data necessary to approximate an efficient frontier in multi-input / multi-output dimensions is greater than 25 and greater than the previously suggested rules of thumb of 8 or 12 recommended in the literature.

This research suggests a small sample multi-input / multi-output efficiency estimation method and a method for identifying outliers and removing them. The results from applying these methods have shown that the outcomes are significantly different if these methods are applied, compared to conventional DEA. Therefore, these methods for

multi-input / multi-output models are recommended when faced with data that may contain outliers or when the data sample is small.

CHAPTER 8

CONCLUSIONS

Data envelopment analysis (DEA) is a system-based approach to performance evaluation that considers multiple inputs and outputs simultaneously. Despite its history and 25 years of application, there are still significant barriers to applying DEA to problems practitioners face. This dissertation attempts to address some of the barriers and to provide alternatives or improvements to the standard DEA technical efficiency analysis. In this chapter, we conclude the thesis by highlighting the contributions in Section 8.1 and some possible future work in Section 8.2.

8.1 Results

Several assumptions are required in using DEA , including:

- The data all represent the same production function and thus are a homogenous group
- The data collected give a good representation of this group
- All data elements are understood and measured correctly

In Chapter 4, an outlier detection methodology is developed to quantify the extent to which these assumptions are true and to identify observations which do not belong in a particular DEA analysis. The super efficiency model is used to quantify a single observation's distance from the rest of the data set. Previously this strategy for identification had only been used to find overly efficient observations. In this work, an inefficient frontier also is constructed in order to identify inefficient observations using a

super inefficiency model. The observations identified as being overly distant from the efficient and inefficient frontiers are flagged for further investigation to understand the source of their difference.

If the number of outliers relative to the data set size is large, this indicates that one of the three previously stated assumptions is violated, and an investigation of the stated assumptions could result in significantly improved modeling and group identification decisions. Thus the outlier detection methodology can be used as a diagnostic tool to quantify the validity of the DEA assumptions for a specific dataset.

Chapter five investigates the orientation decisions for super efficiency models. As noted in chapter four, super efficiency models are very important to the outlier detection. However, they also have an important role in the calculation of Malmquist productivity index. It has been shown previously that there are cases for the input or the output orientation where the associated linear programs are infeasible. In the context of outlier detection this implies cases in which the decision to flag an observation would be made with little or no information. As a remedy to this problem the hyperbolic oriented super efficiency model is suggested and investigated. It is shown that the hyperbolic oriented measure has a feasible solution when all data are positive. Further it is shown, under a relatively weak condition, that if an observation in the reference set has zero values for at least the same set of inputs as the observation under evaluation, the hyperbolic oriented measure can always be calculated. These two results illustrate the benefits of the hyperbolic orientation in super efficiency models.

The efficiency estimates produced by DEA can be demonstrated to be statistically consistent for a wide range of sampling distributions (e.g. Banker [1993], Kneip, Simar

and Wilson [1998], and Gijbels et al. [1999]). Unfortunately, the rate of convergence is low, especially if the number of input-output variables is high. However, many of the applications of DEA in the academic literature involve small samples, with the explanations such as a lack of homogeneous reference units or the proprietary nature of the data making data hard to acquire. With this problem in mind, Chapter 6 develops a multi-input / multi-output quantile based approach (MQBA).

This method uses a distance function to aggregate inputs (or outputs) and then selects a single input (or output) as a basis of normalization. This creates an estimatable equation to which the quantile method can be applied. The results of the method are an efficiency ordering, using the entire range from 0 to 1. A t-test can then be used to determine if two units' efficiency levels are significantly different. While DEA calculates an efficiency estimate based on the distance of an observation from the estimated efficient frontier; MQBA makes no attempt to identify the efficient frontier which is usually impossible in small samples as illustrated by the low rate of convergence. However, by imposing a functional form, an efficiency ordering can be constructed providing information similar to the results of DEA. Thus MQBA is an alternative to DEA when the sample size is small.

The purpose of using DEA is to quantify the performance of a group of enterprises. The ability to distinguish between top performing units and poorer performing units is important for a variety of reasons. In a competitive environment it allows poor performers to understand the quality of their performance and to apply benchmarking techniques to guide them toward improvement. For the smaller set of industry specific data by using the outlier detection methodology, we find that 25 observations are not

enough to support a multi-input/multi-output DEA model. Because we believe this process is truly a multi-input/multi-output, the MQBA efficiency evaluation method is used. The benchmarking results are beneficial to the individual warehouses. However, to the extent that these 25 warehouses characterize the entire industry, the results also give the industry as a whole guidance in making operating decisions and identifying desirable attributes.

In summary the main contribution of this thesis is to introduce several methodological advances:

- the MQBA method allowing efficiency measurement for multi-input / multi-output models in the case of small data sets
- an outlier detection methodology which not only identifies outliers that are overly efficient, but also outliers that are overly inefficient
- a demonstration that the hyperbolic oriented efficiency measure provides several benefits relative to the more common input or output orientations for use in super efficiency models which are used in outlier detection
- a demonstration that results differ significantly when these new methods are used via a case study of warehouse performance

8.2 Future Research

In this area of research there are still significant challenges and questions to answer. Technical efficiency is still a broadly used measure in productivity assessment. A problem with technical efficiency is that it requires the assumption that all points on the efficient frontier are equally efficient. However, when a producer does not value each output or each input equally, this assumption is not correct.

When it is assumed all points on the efficient frontier are equally desirable, and that all inputs are equally valued and all outputs are equally valued, the best way to become efficient is to reduce that input which requires the smallest percentage decrease to reach the efficient frontier or to increase that output which currently achieves the largest percentage of benchmark unit's output value. Quite often this is the hardest input (output) to decrease (increase) because any further change requires using a more extreme input or output mix.

Using technical efficiency alone to drive change is probably not wise; however, its role as a component of economic efficiency is still interesting. In order to calculate economic efficiency, price information is necessary. Thus additional data that was assumed not to be available in this study would be needed. A concern is that price information is often estimated, and for non-commodity goods, each seller may actually have their own price. Thus for studies that wish to use economic efficiency, this uncertainty or variability in prices needs to be modeled. Possible directions for future research would more clearly describe why technical efficiency can be a misleading measure, and would suggest methods for modeling uncertainty or variability in prices.

The case study examining warehouses was primarily concerned with quantifying the performance of warehouse managers. Based on this goal, orientation decisions were made, and boundaries for the system to be modeled were defined. An assumption was the output requirements of the warehouse are generated externally. The warehouse manager tries to minimize inputs relative to given requirements and an input oriented model represents this goal. However, not all input levels can be adjusted with the same frequency meaning some decisions need to be made before the output requirement is

known with certainty. Thus there is a forecasting problem. The warehouse manager plans input levels based on a forecasted output demand. If this forecast is incorrect, and a different output level is realized, it is likely that the manager could have made a better decision, had he/she been given the correct output demand. Thus separating the effects of forecasting error from inefficient operations is still an open question. Related to this issue, how to properly model inputs that cannot be adjusted with the same frequency is also an area for future research. The current model used in DEA assumes that all inputs are completely flexible.

In this study, efficiency has been assumed to be a direct measure of performance. Thus failure to reach the efficient frontier indicates poor performance relative to other observations. The rational unit, once realizing its performance could be improved, would attempt to move towards the efficient frontier. However, in some cases it might cost the unit more in resources expended to become efficient, than the benefits of the gains in efficiency. The idea that some level of inefficiency is rational behavior was introduced in Bogetoft and Hougaard [2003]. Perhaps the most obvious example relates back to the case in which inputs levels are determined based on forecasts, and if these forecasts are exactly correct, a perfectly efficient input set of resources can be used. However, should the forecast be slightly higher or slightly lower, the set of resources used will not be efficient. Because the benefit of being able to capture the unexpected demand beyond the forecasted level often out weights the cost of being inefficient and carrying slightly more resources, rational behavior may explain a certain level of inefficiency. This is a closely related issue to the question of how to properly model inputs that cannot be adjusted

continuously. While Bogetoft and Hougaard [2003] present some simple models, further research in this area would greatly improve DEA's modeling of actual behavior.

Chapter 6 discusses orientation choices in super efficiency models. However, orientation choices in general DEA models are still an area in which DEA relies on relatively strong assumptions. As stated in Chambers, Chung and Fare [1996], the directional distance function corresponds to taking any orientation in a DEA model. However, currently there is no clear way to select a particular direction. There are three orientations that are often chosen: An input orientation which assumes that outputs cannot be adjusted and inputs are completely flexible, an output orientation which assumes that inputs cannot be adjusted and outputs are completely flexible, or a hyperbolic measure which assumes both inputs and outputs are completely flexible. In truth rarely are all inputs and outputs flexible because all inputs and outputs cannot be adjusted with the same frequency. Thus one could build a model with a set of inputs and outputs with variable adjustment frequencies and select an orientation that reflects the different levels of adjustability. What data and how to use this data to select an orientation are an open questions that deserves further research.

Finally a better understanding of the distribution of DEA efficiency estimates and the underlying causes based on the assumptions in the DEA model deserve further investigation. As noted in Chapter 7, a distribution with a large spike of units observed to be efficient, combined with a distribution of the other observations with a mean near 0.5 is a commonly observed result of DEA analysis, Kittelsen [1999]. Other efficiency measurement methods such as stochastic frontier approach (SFA) adopt distribution assumptions closer to what economic theory would suggest, such as exponential or half-

normal distribution. The difference between the distribution of DEA and SFA efficiency estimates is the lack of observations with efficiency on the range of 0.6 to 0.99 in DEA.

A characteristic of DEA related to the distribution results is the presence and in some cases the predominance of anchor points. As defined by Bougnol [2001], an observation j is an anchor point of a production possibility set if and only if it is an observation on the efficient frontier and there exists a supporting hyperplane $H(\hat{u},\hat{v})$ of the production possibility set containing j such that at least one of the coefficients $\hat{u}_{ri} = 0; i = 1,...,s$ or $\hat{v}_i = 0; i = 1,...,m$, where $\hat{u}_{ri}$ and $\hat{v}_i$ are the decision variables in the multiplier version of the DEA linear program. Dula and Bougnol [2005] noted, for models with the sum of inputs and outputs greater than 4, it is very difficult to find an extreme efficient observation that is not an anchor point. An investigation of anchor points and their role in shaping the distribution of efficiency estimates in DEA is another valuable future research.

While there are many problems still to be solved related to performance measurement, this is a very valuable area of research. In order to compare alternative design or operational decision a method of performance measurement needs to be defined. Models to quantify performance will need to be tailored for each type of enterprise. There are many fundamental questions related to performance measurement that have not been answered. It is these that provide interesting future research in this field going forward.

REFERENCES

Afriat, S. N. (1972). "Efficiency estimation of production functions." International Economic Review **13**(3): 568-598.

Aigner, D. and S. Chu (1968). "On estimating the industry production function." American Economic Review **58**: 826-839.

Aigner, D., C. A. K. Lovell and P. Schmidt (1977). "Formulation and estimation of stochastic frontier production function models." Journal of Econometrics **6**: 21-37.

Anderson, P. and N. C. Petersen (1993). "A procedure for ranking efficient units in data envelopment analysis." Management Science **39**(10): 1261-1264.

Anderson, T. and K. B. Hollingsworth (1996). Estimating the number of DMUs needed for DEA as a function of model size. Portland, Portland State University**:** 16.

Banker, R. D. (1984). "Estimating most productive scale size using data envelopment analysis." European Journal of Operational Research **17**: 35-44.

Banker, R. D. (1993). "Maximum-likelihood, consistency and data envelopment analysis - a statistical foundation." Management Science **39**(10): 1265-1273.

Banker, R. D., A. Charnes and W. W. Cooper (1984). "Some models for estimating technical and scale inefficiencies in data envelopment analysis." Management Science **30**: 1078-1092.

Banker, R. D., A. Charnes, W. W. Cooper, J. Swarts and D. Thomas (1989). "An introduction to data envelopment analysis with some of its models and their uses". Research in Governmental and Nonprofit Accounting. Greenwich, JAI Press. **5:** 125-163.

Banker, R. D., R. F. Conrad and R. P. Strauss (1986). "A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production." Management Science **32**(1): 30-44.

Banker, R. D., S. Das and S. M. Datar (1989). "Analysis of cost variances for management control in hospitals". Research in Governmental and Nonprofit Accounting. Greenwich, JAI Press. **5**.

Banker, R. D. and R. C. Morey (1986a). "Efficiency analysis for exogenously fixed inputs and outputs." Operations Research **34**: 513-521.

Banker, R. D. and R. C. Morey (1986b). "The use of categorical variables in data envelopment analysis." Management Science **32**(12): 1613-1627.

Barnett, V. and T. Lewis (1995). Outliers in Statistical Data. Chichester, Wiley.

Bauer, P. W., A. N. Berger, G. D. Ferrier and D. B. Humphrey (1998). "Consistency conditions for regulatory analysis of financial institutions: a comparison of frontier efficiency methods." Journal of Economics and Business **50**(2): 85-114.

Beckman, R. J. and H. J. Trussell (1974). "Distribution of an arbitrary studentized residual and effects of updating in multi-regression." Journal of American Statistical Association **69**(345): 199-201.

Berger, A. N. and D. B. Humphrey (1991). "The dominance of inefficiencies over scale and product mix economies in banking." Journal of Monetary Economics **28**(1): 117-148.

Bogetoft, P. and J. L. Hougaard (2003). "Rational inefficiencies." Journal of Productivity Analysis **20**(3): 243-271.

Bojanic, A. N., S. B. Caudill and J. M. Ford (1998). "Small-sample properties of ML, COLS and DEA estimators of frontier models in the presence of heteroskedasticity." European Journal of Operational Research **108**: 140-148.

Boles, J. N. (1966). Efficiency squared- efficient computation of efficiency indexes. Western Farm Economic Association, Annual Meeting, Los Angles, Pullman.

Boles, J. N. (1971). The 1130 Farrell efficiency system - multiple products, multiple factors, Giannini Foundation of Agricultural Economics, University of California, Berkeley.

Bougnol, M.-L. (2001). Nonparametric frontier analysis with multiple constituencies, University of Mississippi**:** 171.

Boussofiane, A., R. G. Dyson and E. Thanassoulis (1991). "Applied data envelopment analysis." European Journal of Operational Research(52): 1-15.

Chambers, R. G., Y. H. Chung and R. Fare (1996). "Benefit and distance functions." Journal of Economic Theory **70**(2): 407-419.

Charnes, A., W. W. Cooper, A. Y. Lewin and L. M. Seiford (1993). Data Envelopment Analysis: Theory, Methods, and Application. New York, Quorom Books.

Charnes, A., W. W. Cooper and E. Rhodes (1978). "Measuring the efficiency of decision making units." European Journal of Operational Research **2**: 429-444.

Charnes, A., W. W. Cooper and E. Rhodes (1981). "Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through." Management Science **27**(6): 668-697.

Charnes, A., W. W. Cooper and T. Sueyoshi (1988). "A goal programming/constrained regression review of the Bell System breakup." Management Science **34**(1): 1-26.

Charnes, A., J. J. Rousseau and J. H. Semple (1996). "Sensitivity and efficiency classifications in the additive model of data envelopment analysis." Journal of Productivity Analysis **7**: 5-18.

Cherchye, L. and T. Post (2003). "Methodological advances in DEA: A survey and an application for the Dutch electricity sector." Statistica Neerlandica **57**(4): 410-438.

Coelli, T. and S. Perelman (1999). "A comparison of parametric and non-parametric distance functions: with application to European railways." European Journal of Operational Research **117**: 326-339.

Coelli, T., D. S. P. Rao and G. E. Battese (1998). An introduction to efficiency and productivity analysis. Boston, Kluwer Academic Publishers.

Cook, R. D. and S. Weisberg (1982). Residuals and Influence in Regression. New York, Chapman and Hall.

Cook, W. D. and J. Zhu (2005). Modeling Performance Measurement : Applications and Implementation Issues in DEA. New York, Springer.

Cooper, W. W., L. M. Seiford and K. Tone (2000). Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software. Boston, Kluwer Academic Publishers.

Cuesta, R. A. and J. L. Zofio (2005). "Hyperbolic efficiency and parametric distance functions: With application to Spanish savings banks." Journal of Productivity Analysis **24**(1): 31-48.

Cummins, J. D. and H. Zi (1998). "Comparison of frontier efficiency methods: An application to the U.S. life insurance industry." Journal of Productivity Analysis **10**(2): 131-152.

Debreu, G. (1951). "The coefficient of resource utilization." Econometrica **19**: 273-292.

Deprins, D., L. Simar and H. Tulkens (1984). "Measring labor inefficiency in post offices". The Performance of Public Enterprises: Concepts and Measurements. M. Machand, P. Pestieau and H. Tulkens. Amsterdam, North-Holland**:** 243-267.

Drake, L. and R. Simper (2003). "The measurement of English and Welsh police force efficiency: A comparison of distance function models." European Journal of Operational Research **147**(165-186).

Dula, J. and M.-L. Bougnol (2005). Role of anchor points in the geometry of DEA. INFORMS Annual Meeting, San Francisco.

Dyson, R. G., R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico and E. A. Shale (2001). "Pitfalls and protocols in DEA." European Journal of Operational Research **132**: 245-259.

Fare, R. and S. Grosskopf (1983). "Measuring congestion in production." Zeitschrift Fur Nationalokonomie-Journal of Economics **43**(3): 257-271.

Fare, R., S. Grosskopf, B. Lindgren and P. Roos (1994). "Productivity developments in Swedish hospitals: A Malmquist output index approach". Data Envelopment Analysis: Theory, Methodology and Applications. A. Charnes, W. W. Cooper, A. Y. Lewin and L. M. Seiford. Boston, Kluwer Academic Publishers.

Fare, R., S. Grosskopf and C. A. K. Lovell (1985). The Measurement of Efficiency of Production. Boston
Hingham, MA, U.S.A., Kluwer-Nijhoff Pub. ;
Distributors for North America, Kluwer Academic Publishers.

Fare, R., S. Grosskopf and C. A. K. Lovell (1994). Production Frontiers. Cambridge [England] ; New York, NY, USA, Cambridge University Press.

Fare, R., S. Grosskopf and O. Zaim (2002). "Hyperbolic efficiency and return to the dollar." European Journal of Operational Research **136**(3): 671-679.

Fare, R. and D. Primont (1995). Multioutput Production and Duality: Theory and Application. Boston, Kluwer Academic Publishers.

Fare, R. and L. Svensson (1980). "Congestion of production factors." Econometrica **48**(7): 1745-1753.

Farrell, M. J. (1957). "The measurement of productive efficiency." Journal of the Royal Statistical Society. Series A. General **120**: 253-281.

Farrell, M. J. (1959). "The Convexity assumption in the theory of competitive markets." Journal of Political Economy **67**(4): 377-391.

Farrell, M. J. and M. Fieldhouse (1962). "Estimationg efficient productions functions under increasing returns to scale." Journal of the Royal Statistical Society. Series A. General **125**: 252-267.

Forsund, F. and N. Sarafoglou (2002). "On the origins of data envelopment analysis." Journal of Productivity Analysis **17**: 23-40.

Fried, H. O., C. A. K. Lovell and S. S. Schmidt (1993). The Measurment of Producitive Efficiency. New York, Oxford University Press.

Fried, H. O., C. A. K. Lovell, S. S. Schmidt and S. Yaisawarng (2002). "Accouting for environmental effects and statistical noise in data envelopment analysis." Journal of Productivity Analysis **17**: 157-174.

Fried, H. O., S. S. Schmidt and S. Yaisawarng (1999). "Incorporating the operating environment into a nonparametric measure of technical efficiency." Journal of Productivity Analysis **12**: 249-267.

Fuentes, H. J., E. Grifell-Tatje and S. Perelman (2001). "A parametric distance function approach for Malmquist productivity index estimation." Journal of Productivity Analysis **15**: 79-94.

Gabrielsen, A. (1975). On estimating efficient production functions. Working Paper, No. A-85. Bergen, Norway, Chr. Michelsen Institute, Department of Humanities and Social Sciences.

Gijbels, I., E. Mammen, B. U. Park and L. Simar (1999). "On estimation of monotone and concave frontier functions." Journal of the American Statistical Association **94**(445): 220-228.

Gong, B.-H. and R. C. Sickles (1992). "Finite sample evidence on the performance of stochastic frontiers and data envelopment analysis using panel data." Journal of Econometrics **51**: 259-284.

Griffin, P. M. and P. H. Kvam (1999). "A quantile-based approach for relative efficiency measurement." Managerial and Decision Economics **20**: 403-410.

Gunst, R. F. and R. L. Mason (1980). Regression Analysis and its Application. New York, Marcel Dekker.

Hackman, S. T., E. H. Frazelle, P. M. Griffin, S. O. Griffin and D. A. Vlasta (2001). "Benchmarking warehousing and distribution operations: An input-output approach." Journal of Productivity Analysis **16**: 79-100.

Hall, M. and C. Winsten (1959). "The ambiguous notion of efficiency." The Economic Journal **69**(273): 71-86.

Hand, D. J. (1981). Discrimination and Classification. Chichester [Eng.] ; New York, Wiley.

Hand, D. J. (1997). Construction and Assessment of Classification Rules. Chichester ; New York, Wiley.

Hjalmarsson, L., S. C. Kumbhakar and A. Heshmati (1996). "DEA, DFA and SFA: A comparison." Journal of Productivity Analysis **7**(2/3): 303-327.

Institute of Industrial Engineers. (2005). "Definition of industrial engineering." from http://www.iienet.org/public/articles/details.cfm?id=468.

Johnson, A. L. and L. F. McGinnis (2005a). An outlier detection methodology with consideration for an inefficient frontier. Working Paper. Georgia Institute of Technology.

Johnson, A. L. and L. F. McGinnis (2005b). What does the distribution of DEA scores look like?, Georgia Institute of Technology. **Working Paper**.

Jondrow, J., C. A. K. Lovell, I. Materov and P. Schmidt (1982). "On the estimation of technical inefficiency in the stochastic frontier production function model." Journal of Econometrics **19**(2/3): 233-238.

Kittelsen, S. A. C. (1999). Monte Carlo simulations of DEA efficiency measures and hypothesis tests. Oslo, University of Oslo**:** 63.

Klein, L. R. (1962). An Introduction to Econometrics. Englewood Cliffs, N.J.,, Prentice-Hall.

Kneip, A. and L. Simar (1996). "A general framework for frontier estimation with panel data." Journal of Productivity Analysis **7**: 187-212.

Kneip, A., L. Simar and P. W. Wilson (1998). "A note on the convergence of nonparametric DEA estimators for production efficiency scores." Econometric Theory **14**: 783-793.

Koopmans, T. C. (1951). "An analysis of production as an efficient combination of activities". Activity Analysis of Production and Allocation. T. C. Koopmans. New York, Wiley. **Monograph No. 13**.

Liu, F.-h. F. and T.-n. D. Hsu (2004). "Least-efficient frontiers of data envelopment analysis-CCR model." Working paper.

Lovell, C. A. K. (1993). "Production frontiers and productive efficiency". The Measurement of Productive Efficiency. H. O. Fried, C. A. K. Lovell and S. S. Schmidt. New York, Oxford University Press. **1:** 3-67.

Lovell, C. A. K. (1994). "Linear programming approaches to the measurement and analysis of productive efficiency." TOP **2**: 175-248.

Lovell, C. A. K., P. Travers, S. Richardson and L. L. Wood (1993). "Resources and functioning: A new view of inequality in Australia". Models and Measurement of Welfare & Inequality. W. Eichorn, Springer Verlag.

Lovell, C. A. K., L. C. Walters and L. L. Wood (1993). "Stratified models of education production using DEA and regression analysis". Data Envelopment Analysis: Theory, Methods, and Application. A. Charnes, W. W. Cooper, A. Y. Lewin and L. M. Seiford. New York, Quorom Books.

McCarty, T. and S. Yaisawarng (1993). "Technical efficiency in New Jersey school districts". The Measurement of Productive Efficiency. H. O. Fried, C. A. K. Lovell and S. S. Schmidt. New York, Oxford University Press. **1:** 271-287.

Mundlak, Y. (1961). "Empirical production function free of management bias." Journal of Farm Economics **43**: 44-56.

Neter, J., W. Wasserman and M. H. Kutner (1985). Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Homewood, Illinois, Richard D. Irwin Inc.

Ondrich, J. and J. Ruggiero (2001). "Efficiency measurement in the stochastic frontier model." European Journal of Operational Research **129**(2): 434-442.

Paradi, J. C., M. Asmild and P. C. Simak (2004). "Using DEA and worst practice DEA in credit risk evaluation." Journal of Productivity Analysis **21**(2): 153-165.

Petersen, N. C. (1990). "Data envelopment analysis on a relaxed set of assumptions." Management Science **36**(3): 305-314.

Quenouille, M. H. (1956). "Notes on bias in estimation." Biometrika **43**(3/4): 353-360.

Ray, S. C. (1988). "Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation." Socio-Economic Planning Science **22**(4): 167-176.

Ray, S. C. (1991). "Resource-use efficiency in public schools: A study of Connecticut data." Management Science **37**(12): 1620-1628.

Rhodes, E. (1978). Data envelopment analysis and related approaches for measuring the efficiency of decision-making units with an application to program Follow Through in U.S. education. School of Urban and Public Affairs. Pittsburgh, Pa, Carnegie-Mellon University.

Rousseeuw, P. J. and B. C. van Zomeren (1990). "Unmasking multivariate outliers and leverage points." Journal of American Statistical Association **85**: 633-639.

Ruggiero, J. (1996). "On the measurement of technical efficiency in the public sector." European Journal of Operational Research **90**: 553-565.

Ruggiero, J. (1998). "Non-discretionary inputs in data envelopment analysis." European Journal of Operational Research **111**: 461-469.

Ruggiero, J. (2004). "Performance evaluation when non-discretionary factors correlate with technical efficiency." European Journal of Operational Research **159**: 250-257.

Schmidt, P. and R. C. Sickles (1984). "Production frontiers and panel data." Journal of Business and Economic Statistics **2**(4): 367-374.

Seiford, L. M. and J. Zhu (1998). "Sensitivity analysis of DEA models for simultaneous changes in all the data." Journal of the Operational Research Society **49**(10): 1060-1071.

Seiford, L. M. and J. Zhu (1999). "Infeasibility of super-efficiency data envelopment analysis models." Infor **37**(2): 174-187.

Shephard, R. W. (1953). Cost and Production Functions. New Jersey, Princeton University Press.

Shephard, R. W. (1970). Theory of Cost and Production Functions. Princeton, N.J., Princeton University Press.

Simar, L. (2003). "Detecting outliers in frontier models: A simple approach." Journal of Productivity Analysis **20**: 391-424.

Simar, L. and P. W. Wilson (1998). "Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models." Management Science **44**(1): 49-61.

Simar, L. and P. W. Wilson (2000). "A gerneral methodology for bootstrapping in non-parametric frontier models." Journal of Applied Statistics **27**(6): 779-802.

Simar, L. and P. W. Wilson (2005). "Estimation and inference in two-stage, semi-parametric models of production processes." Journal of Econometrics **Forthcoming**.

Thore, S., G. Kozmetsky and F. Phillips (1994). "DEA of financial statements data: the U.S. computer industry." Journal of Productivity Analysis **5**: 229-248.

Thrall, R. M. (1996). "Duality, classification and slacks in DEA." Annals of Operations Research **66**: 109-138.

Timmer, C. P. (1971). "Using a probabilistic frontier production function to measure technical efficiency." Journal of Political Economy **79**: 767-794.

Tukey, J. W. (1958). "Bias and confidence in not-quite large samples." Annals of Mathematical Statistics **29**(2): 614.

Tulkens, H. and P. Vanden Eeckaut (1995). "Non-parametric efficiency, progress and regress measures for panel data: Methodological aspects." European Journal of Operational Research **80**: 474-499.

Wilson, P. W. (1995). "Detecting influential observations in data envelopment analysis." Journal of Productivity Analysis **6**: 27-45.

Yu, C. (1998). "The effect of exogenous variables in efficiency measurement - A Monte Carlo study." European Journal of Operational Research **105**: 569-580.

Zhu, J. (1996). "Robustness of the efficient DMUs in data envelopment analysis." European Journal of Operational Research **90**(3): 451-460.

Zofio, J. L. and C. A. K. Lovell (2001). "Graph efficiency and productivity measures: an application to US agriculture." Applied Economics **33**(11): 1433-1442.

VITA

ANDREW L. JOHNSON

Andy was born to Leslie G. Johnson Jr. and Rebecca Musser Johnson on July 10th, 1978 in Warrenton, Virginia. He attended public school in Fairfax Virginia and received a B.S. in Industrial and Systems Engineering from Virginia Polytechnic Institute and State University, Blacksburg , Virginia in 2001. As an undergraduate, he had the opportunity to study abroad at Tohoku University in Sendai, Japan and work as a co-op student for Ingersoll-Rand in Roanoke, Virginia and in Yokohama, Japan. He completed a Master's of Science in Industrial and Systems Engineering in 2002 and a Doctor of Philosophy of Industrial and Systems Engineering in 2006. He is married to Chiaki Kajiro and when he is not working on his research, enjoys traveling and walking with his dog, Allie.